Supplementary Material
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

In section 1 we report ablation results on episodic memory tokens and context consistency loss. Section 2 provides insights into different transformer-based encoders adopted in our approach. In section 3 we report additional results of baselines over different data sources, evaluation of our approach on other datasets, and comparison of maneuver classes. Section 4 shows gaze fixation over time, followed by sensor details in section 5 and limitations and future scope in section 6.

1 Ablation Study

In Table 1, we observe that there is an increment in accuracy and F_1 score when the input embeddings are augmented with learnable memory tokens and the context-consistency loss function is taken into consideration. Specifically, we observe a 3.56% and 4.58% increase in accuracy and F_1 score on M^{2}MVT, 32×3, suggesting an improvised performance in the presence of both EM and CCL. Moreover, M^{2}MVT, 32×3 achieves a 1.36% higher accuracy than M^{2}MVT, 16×4, pointing to a higher performance due to longer sampling.

2 Additional Details of Transformer Encoders

Fig. 2 shows the detailed architecture of the different transformer-based encoders used during baselining. MViT [2] follows a multiscale feature hierarchy where the model learns by progressively decreasing the spatial resolution while increasing the channel capacity. On top of it, MViTv2 [5] takes the decomposed relative positional embeddings and a residual pooling connection, in contrast to the joint patch and positional embeddings for the ViT-based encoders.

In all these methods, we maintain consistency in utilizing pre-training, fine-tuning, and spatio-temporal patching strategies. The same encoder is used for processing the different modalities due to its impressive performance on relevant tasks. All of these networks are augmented with $N = 12$ learnable memory embeddings. We evaluate them both with and without the memory tokens and context-consistency loss, and summarize the results in Table 1.

Table 1: Performance comparison of ViT-based CEMFormer [6], MViT [2], MViTv2 [5], and the proposed M^{2}MVT (16×4 and 32×3) with and without Episodic Memory Embeddings (EM) [9] and Context Consistency Loss (CCL) [6] on our dataset. In cases without CCL, standard cross-entropy loss was used.

Encoder	Module EM	Module CCL	Acc. (↑)	F_1 (↑)
ViT [6]	✗	✗	57.19	58.55
	✗	✓	58.25	59.10
	✓	✗	60.05	62.21
	✓	✓	60.74	63.09
MViT [2]	✗	✗	58.73	60.23
	✗	✓	59.19	60.48
	✓	✗	61.04	62.10
	✓	✓	62.13	63.65
MViTv2 [5]	✗	✗	59.31	60.37
	✗	✓	60.68	61.03
	✓	✗	61.26	62.32
	✓	✓	62.78	64.08
M^{2}MVT (16×4)	✗	✗	60.67	60.43
	✗	✓	62.24	63.62
	✓	✗	62.97	64.15
	✓	✓	64.23	65.11
M^{2}MVT (32×3)	✗	✗	62.03	62.14
	✗	✓	63.69	64.97
	✓	✗	64.17	65.03
	✓	✓	**65.59**	**66.72**

3 Additional Experiments

This section gives an overview of the supplementary experiments done for early anticipation of driving maneuvers.

3.1 Evaluation of Baselines over Different Data Sources

Table 2 serves as the motivation for shifting from recurrent networks like LSTMs, to the more recent vision transformer-based architectures. This is essentially a precursor to Table 4 in the main paper which has a comparison to the proposed approach. The LSTM-based models [3, 8] exhibit a lower accuracy and F_1 score with 3 to 5 times more parameters than the ViT-based CEMFormer. Moreover, their parameters surge at a higher rate than [6], every time a view is added. Thus, LSTM-based models are computationally prohibitive for multi-view multi-modal data captured over longer durations. It is important to note that the performance of all the baselines improves when the model is trained on gaze maps alongside the ego view (column 4).

3.2 Evaluation of M^{2}MVT on Other Datasets

In Table 3, we observe that M^{2}MVT achieves higher accuracy than the other baselines on the given datasets. On Brain4Cars, M^{2}MVT outperforms the state-of-the-art CEMFormer by 1.38% on F_1 for both in- and out-cabin views. For

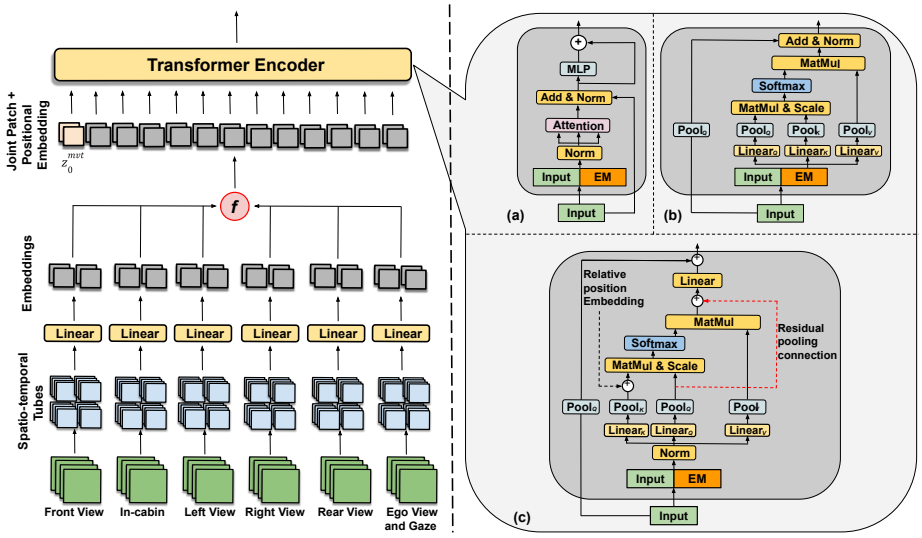

Fig. 1: Different transformer-based encoders used during baselining. (a) Standard ViT [6] (b) Multiscale Vision transformer [2] with episodic memory (c) MViTv2 [5]. Here EM are the learnable episodic memory embeddings.

Table 2: Comparison of accuracy, F_1 score (%), and Parameters (in Millions) on various methods like Gebert *et. al* [3], Rong *et. al* [8], and CEMFormer [6]

.

Method	Metric	Data Source						
		Aria RGB	Aria Gaze	In-Cabin & Aria RGB	In-Cabin & Gaze	Out-Cabin & Aria RGB	Out-Cabin & Gaze	All Views & Gaze
Gebert *et. al* [3]	**Acc.(↑)**	32.94	35.52	41.69	44.72	53.01	53.17	54.03
	F_1 (↑)	30.67	33.19	43.08	45.82	54.16	54.65	54.91
	Param (M) (↓)	240.26	240.26	325.52	325.52	496.04	496.04	581.73
Rong *et. al* [8]	**Acc.(↑)**	33.71	36.12	41.96	44.37	51.85	52.13	54.89
	F_1 (↑)	33.19	35.64	43.84	47.42	54.44	54.13	55.01
	Param (M) (↓)	46.22	46.22	212.92	212.92	286.83	286.83	324.62
CEMFormer [6]	**Acc.(↑)**	37.17	42.02	47.78	51.33	53.80	59.12	**60.74**
	F_1 (↑)	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**

datasets like HDD and VIENA^2, which only capture the out-cabin videos, M^{2}MVT achieves 4.11% and 2.11% higher accuracy respectively. Furthermore, its accuracy is 3.41% higher than CEMFormer on our dataset. This performance can be attributed to the robust temporal modeling exhibited by M^{2}MVT, a characteristic not observed in conventional vision transformers. Moreover, its incorporation of spatiotemporal patching and pooling, alongside adaptable memory tokens, and decomposed relative positional embeddings, contributes to enhanced performance in maneuver anticipation tasks.

Table 3: Performance comparison of M^{2}MVT. Accuracy and F_1 score (%) over different data sources on the Brain4Cars [4], HDD [7], VIENA^2 [1], and our dataset[3]. "-" indicates that the datasets do not have an in-cabin data source.

Method	Data Source	Brain4Cars Acc.(↑)	Brain4Cars F_1(↑)	HDD Acc.(↑)	HDD F_1(↑)	VIENA^2 Acc.(↑)	VIENA^2 F_1(↑)	DAAD (Ours) Acc.(↑)	DAAD (Ours) F_1(↑)
M^{2}MVT	**In-Cabin**	81.87	80.90	-	-	-	-	50.43	48.11
	Out-Cabin	64.07	65.35	72.51	71.89	75.63	73.47	58.78	59.91
	In and Out-Cabin	83.18	84.11	-	-	-	-	61.74	63.82

Table 4: Performance comparison of baselines and M^{2}MVT accuracy and F_1 score (%) for all the maneuver classes of Brain4Cars, HDD, VIENA^2, AIDE and our dataset. Note: AIDE dataset [10] was unavailable at the time these experiments were being conducted on M^{2}MVT and was reported to be corrupted on the official github repo.

Dataset	Class	Method: Gebert *et. al* [3] Acc. (↑)	Gebert *et. al* [3] F_1 (↑)	Rong *et. al* [8] Acc. (↑)	Rong *et. al* [8] F_1 (↑)	CEMFormer [6] Acc. (↑)	CEMFormer [6] F_1 (↑)	M^{2}MVT Acc. (↑)	M^{2}MVT F1. (↑)
Brain4Cars [4]	ST	80.36	79.14	84.68	82.48	87.36	85.73	90.07	89.34
	RT	75.06	70.71	78.93	72.39	82.13	77.18	84.42	85.61
	LT	74.59	70.89	79.62	73.27	78.67	72.49	83.05	82.11
	RLC	70.38	65.57	71.43	64.09	75.31	69.46	81.49	81.73
	LLC	67.44	62.04	73.89	64.93	76.05	70.16	82.19	80.94
HDD [7]	ST	69.57	71.49	71.28	73.20	75.87	74.67	78.37	76.49
	RT	51.15	49.37	56.82	59.23	70.06	72.30	73.08	72.16
	LT	59.88	54.28	61.64	63.82	68.58	71.62	71.38	70.21
	RLC	42.71	47.38	45.61	49.27	59.15	63.57	63.73	63.10
	LLC	44.53	46.37	47.49	49.20	63.94	65.02	67.39	70.02
	SS	66.93	67.04	68.95	70.23	74.63	70.14	75.21	74.45
	UT	38.33	30.12	40.19	34.60	44.74	39.25	49.76	50.15
VIENA^2 [1]	ST	71.81	72.07	78.21	77.23	80.45	78.17	83.42	85.17
	RT	64.59	65.46	69.44	65.71	71.70	73.15	75.23	75.67
	LT	65.67	62.79	69.26	67.65	73.39	74.94	77.29	76.20
	RLC	61.74	59.96	65.92	66.72	67.95	70.08	71.18	70.53
	LLC	59.82	56.56	62.18	59.36	64.62	69.38	68.26	71.59
	SS	73.53	69.35	77.26	74.60	82.53	83.90	85.41	85.94
AIDE [10]	GS	69.57	71.49	71.28	73.20	75.87	74.67	-	-
	RT	51.15	49.37	56.82	59.23	70.06	72.30	-	-
	LT	59.88	54.28	61.64	63.82	68.58	71.62	-	-
	RLC	42.71	47.38	45.61	49.27	59.15	63.57	-	-
	LLC	44.53	46.37	47.49	49.20	63.94	65.02	-	-
DAAD (Ours)	ST	55.29	56.13	58.11	57.10	64.87	66.49	67.36	68.73
	RT	47.89	49.20	49.36	44.39	63.29	60.08	66.48	65.14
	LT	45.71	47.28	46.37	49.25	59.41	57.90	67.29	69.71
	RLC	38.02	36.88	43.28	39.17	47.16	50.18	53.51	55.37
	LLC	38.63	33.65	38.92	34.54	48.09	51.22	55.25	55.23
	SS	51.19	50.28	56.73	57.89	60.92	59.17	67.94	65.35
	UT	19.27	14.72	21.64	22.68	31.26	25.63	47.51	45.06

3.3 Comparison of Maneuver Classes on Baselines

Table 4 gives the per-class accuracy and F_1 scores across the benchmarked datasets. We infer that U-turn is the most complicated maneuver to be predicted, owing to its low prediction score on HDD and DAAD. Also, although both HDD and VIENA^2 are captured using only a front-facing camera, accuracy and F_1 scores are higher on VIENA^2. It can be attributed to two reasons: Firstly, VIENA^2 is captured in a simulated setup with many identical maneuvers. Real-

world driving is more complex than a simulator can encompass. Secondly, HDD maneuver clips have a longer duration than VIENA^2, indicating that temporal modeling over longer durations is a more complex task. Moreover, M^{2}MVT outperforms other methods on accuracy and F_1 score across maneuvers.

4 Gaze fixations over time

In Fig. 2, we plot the variation of gaze over time, before the onset of a maneuver. In Fig. 2(a), the driver is about to take a left turn. As observed, the driver's gaze gradually starts to move tangential towards the turn, as is generally observed in cases where the driver is about to take a turn. A similar pattern is observed in Fig. 2(b), however towards a right turn. We infer that gaze from the egocentric view provides strong cues for early maneuver anticipation, which is further validated by the results from Table 2, and main paper (see Table 4).

Fig. 2: Gaze fixations before the onset of a maneuver (Left to Right) (a) Left Turn. (b) Right Turn.

5 Additional Details of Sensors

Camera configurations for the data capture:

In-cabin

– 1×GoPro 8 camera mounted inside the car with driver-facing view while driving. Resolution is 1080p and a frame rate of 30 fps, horizontal field of view is 85.8° and vertical field of view is 55.2°, hypersmooth stabilization to be set on and an aspect ratio of 16 : 9.

- 1×OBD device with GPS speed (meters/second), acceleration sensor, latitude, longitude, G sensor calibration, as well as speed and throttle enabled.
- 1×Meta Aria, a wearable glass that captures ego-centric videos. We choose profile 15 from the set of profiles enabled on Aria, that has eye-tracking cameras (resolution: 320×240, frame rate: 30 fps), RGB cameras (resolution: 1408×1408, frame rate: 30 fps), Mono scene cameras (resolution: 640×480, frame rate: 30 fps), SLAM, IMU's, and a microphone.

Out-cabin

- 4×GoPro 8 cameras mounted on the front, rear, and side-mirror views (left, right) of the car capturing the surrounding traffic contextual situation while driving. Resolution is 1080p and frame rate of 30 fps, field of views (except for the front camera): 85.8° for horizontal and 55.2° for vertical, hypersmooth stabilization to be set on and an aspect ratio of 16 : 9. While the horizontal and vertical field of views for the front-facing camera are 118.2° and 69.5° respectively.

6 Limitations and Future Scope

DAAD has been majorly captured in Indian road scenarios. So, it would be challenging to generalize well across geographies having different traffic and driving scenarios. Therefore, jointly training the model on several datasets can be a future direction so that it accommodates diverse scenarios and adapt better to various geographies. Moreover, annotations for cause, dynamic agents, and static agents could benefit future work in this direction. Finaly, the model can be improvised to accommodate for the poor performance from the analysis of the confusion matrix.

References

1. Aliakbarian, M.S., Saleh, F.S., Salzmann, M., Fernando, B., Petersson, L., Andersson, L.: Viena2: A driving anticipation dataset (2018)
2. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: ICCV. pp. 6824–6835 (2021)
3. 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)
4. 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)
5. 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)
6. 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)

7. 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)
8. 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)
9. Sandler, M., Zhmoginov, A., Vladymyrov, M., Jackson, A.: Fine-tuning image transformers using learnable memory. In: CVPR. pp. 12155–12164 (2022)
10. 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)