Towards Safer and Understandable Driver Intention Prediction

Supplementary Material

1. Ambiguity Analysis on DAAD-X

We employed 10 annotators with prior experience in both driving and annotation. All annotators were specifically trained to focus on key driving-related visual cues (e.g., maneuver window, surrounding vehicles, lane structure) to ensure consistency. We define the *intention window* as the interval between the annotators' marked start and end points of a maneuver, capturing the temporal span from when the driver's intention becomes observable to when the action completes.

To assess annotation quality, we measured inter–annotator agreement using vote variance and found an overall ambiguity rate of less than 5%, as described in Table 1. Ambiguous cases were resolved through majority voting (see Sec. 3.1 of the main paper). The average ambiguity across the dataset was 4.04%, which was further reduced through refinement. Initially, $1{,}725$ videos were annotated, after removing ambiguous cases or corrupted samples, the final dataset comprised of $1{,}568$ videos.

Table 1. Intention labels with respect to maneuver ambiguity analysis. This shows the $\%$ of videos which has ambiguous annotations. The variance between votes across annotators differs a lot for ambiguous cases.

Metric	ST	SS	LT	LLC	RT	RLC	UT
# Videos	438	229	341	192	234	237	54
Ambiguity (%)	6.40	4.80	2.67	6.93	0.00	1.06	6.40

2. Explanation Annotation using VLMs

We initially explored using vision-language models (VLMs) (e.g., InternVideo-2 [3]) to automate concept annotation for each video, aiming to reduce manual annotation effort. However, as illustrated in Fig. 1, the results were suboptimal. Existing publicly available VLMs are typically trained to generate generic scene descriptions and are not fine-tuned for driving-specific tasks. Consequently, the extracted concepts were often too broad or irrelevant, leading to ambiguity in neuron activations within the bottleneck layer.

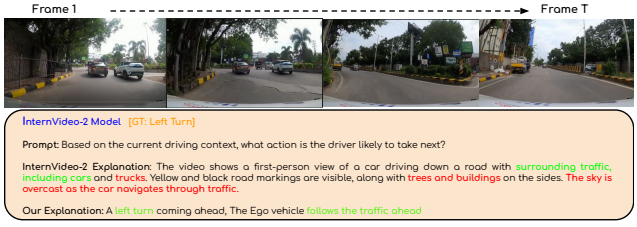

Figure 1. Response of the InternVideo-2 [3] 8B model to a Left Turn video.

Table 2. Distribution for Eye-Gaze Explanations. This shows the number of instances for each of the eye gaze explanations.

Label Index	Explanations	Instances
1	The Gaze is mostly towards the forward direction	223
2	The Gaze is mostly towards the traffic ahead	65
3	The Gaze is towards the action inducing objects	147
4	The Gaze is towards the right side mirror	19
5	The Gaze is on the left side and slowly changes to right side	20
6	The Gaze is mostly towards the front right side	84
7	The Gaze is initially towards left side and then moves to the right side	21
8	The Gaze is mostly towards the right side	29
9	The Gaze is mostly towards the right side mirror	117
10	The Gaze is towards the front right side	26
11	The Gaze is mostly towards the front left side	151
12	The Gaze is initially towards the right side and then moves to left side	64
13	The Gaze is mostly to the left side	10
14	The Gaze is mostly towards the left side mirror	96
15	The Gaze is towards the front left side	25

Table 3. Distribution for Ego Vehicle Explanations. This shows the number of instances for each of the ego vehicle explanations.

Label Index	Explanations	Instances
1	The Ego vehicle is nearing an intersection and there's no traffic light	70
2	The Ego vehicle is nearing an intersection and traffic light is green	15
3	The Ego vehicle follows the traffic ahead	360
4	The road is clear ahead	129
5	The Ego vehicle deviates to avoid slow vehicle/obstacles and moves straight	31
6	The vehicle ahead slows down	114
7	Pedestrian or Vehicle cutting in the path ahead	45
8	The Ego vehicle is nearing an intersection and traffic light is red	23
9	The Ego vehicle joins the traffic moving to the right	29
10	A right turn coming ahead	118
11	The traffic moves from the left to right side at the intersection	44
12	The traffic moves from the right to left side at the intersection	55
13	The road is clear ahead on the right lane	100
14	No speeding vehicle on the right lane is coming from the rear right side	12
15	A left turn coming ahead	200
16	The road is clear ahead on the left lane	84
17	No speeding vehicle on the left lane is coming from the rear left side	10

3. DAAD-X Dataset Explanations

Table 2 presents the number of instances corresponding to each gaze-based explanation along with their associated actions (maneuvers). Similarly, Table 3 lists the instance counts for each ego-vehicle explanation and its related actions. The mapping of anchor labels used in the multi-label t-SNE plot (see Fig. 8 of the main paper) can be referenced from the Ego Explanation Table 3.

4. Implementation Details

Data Augmentations. We sample 16 frames from each video and convert them into tubelets using Conv3D for transformers and patches from Conv2D for CNN architectures. To mitigate noise and overfitting, we apply the following data augmentations: (1) Flip frames along the Y-axis to transform left maneuvers into right maneuvers, and (2) Divide every video into 16 equal frame segments, randomly selecting one per segment to introduce random temporal shuffling. For gaze-overlaid videos, we model gaze coordinates using a Gaussian distribution with a standard deviation of 100 in pixel

space. For gaze cropping experiments, we use a variable radius r in pixel space, where $r \in \{50{,}150{,}250{,}350{,}450{,}550\}$.

Training Details. All experiments were conducted on a single RTX A6000 48GB GPU with a fixed batch size of 8. We used the AdamW optimizer with a weight decay of 0.05, a base learning rate of 5 x 10^{-5}, and a dropout rate of 0.45. Following previous works [1, 2], we finetune the attention layers of the backbone, the final MLP classifiers, and our bottleneck blocks. VideoMAE and MViTv2 are trained for 200 epochs, while I3D is trained for 100 epochs. Furthermore, we apply a cosine annealing learning rate scheduler with warm restarts, starting at 50 epochs with a reset multiplier of 2 for the VideoMAE and MViTv2 models.

Evaluation Metrics. We evaluated our model on two tasks: (1) a multiclass classification task for maneuver prediction and (2) a multi-label classification task for ego-vehicle explanations, using accuracy and F_1 score as performance metrics. For maneuver classification, accuracy is the ratio of correctly predicted instances to total instances, while the F_1 score is the harmonic mean of precision and recall. For multi-label classification task, accuracy is defined as the proportion of instances where all predicted labels match the ground truth, and the F_1 score is computed as a weighted average of per-class F_1 scores, including both macro and micro averaging. These metrics help address class imbalance and ensure a balance between precision and recall across both tasks.

References

[1] Yunsheng Ma, Wenqian Ye, Xu Cao, Amr Abdelraouf, Kyungtae Han, Rohit Gupta, and Ziran Wang. Cemformer: Learning to predict driver intentions from in-cabin and external cameras via spatial-temporal transformers. In *IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 4960–4966, 2023.

[2] Koen Vellenga, H Joe Steinhauer, Göran Falkman, and Tomas Björklund. Evaluation of video masked autoencoders' performance and uncertainty estimations for driver action and intention recognition. In *WACV*, pages 7429–7437, 2024.

[3] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, and et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV (85)*, pages 396–416. Springer, 2024.