Towards Safer and Understandable Driver Intention Prediction

Mukilan Karuppasamy[1] Shankar Gangisetty[1] Shyam Nandan Rai[2] Carlo Masone[2] C V Jawahar[1]
[1]*IIIT Hyderabad, India* [2]*Politecnico di Torino, Italy*

Abstract

*Autonomous driving (AD) systems are becoming increasingly capable of handling complex tasks, mainly due to recent advances in deep learning and AI. As interactions between autonomous systems and humans increase, the interpretability of decision-making processes in driving systems becomes increasingly crucial for ensuring safe driving operations. Successful human-machine interaction requires understanding the underlying representations of the environment and the driving task, which remains a significant challenge in deep learning-based systems. To address this, we introduce the task of interpretability in maneuver prediction before they occur for driver safety, i.e., driver intent prediction (DIP), which plays a critical role in AD systems. To foster research in interpretable DIP, we curate the eXplainable Driving Action Anticipation Dataset (**DAAD-X**), a new multimodal, ego-centric video dataset to provide hierarchical, high-level textual explanations as causal reasoning for the driver's decisions. These explanations are derived from both the driver's eye-gaze and the ego-vehicle's perspective. Next, we propose Video Concept Bottleneck Model (**VCBM**), a framework that generates spatio-temporally coherent explanations inherently, without relying on post-hoc techniques. Finally, through extensive evaluations of the proposed VCBM on the DAAD-X dataset, we demonstrate that transformer-based models exhibit greater interpretability than conventional CNN-based models. Additionally, we introduce a multilabel t-SNE visualization technique to illustrate the disentanglement and causal correlation among multiple explanations. Our data, code and models are available at:* `https://mukil07.github.io/VCBM.github.io/`

1. Introduction

The increasing reliance on deep neural networks in safety-critical applications [26] raises significant concerns due to their black-box nature, resulting in a lack of interpretability. In autonomous driving [16, 38], this lack of transparency makes it difficult for users to trust AI-driven decisions, leading to safety and accountability challenges, particularly in accidents. Ensuring safer deployment requires models that predict driving actions and provide human-understandable explanations for their decisions.

For instance, consider the scenario illustrated in Fig. 1. An autonomous car is traveling at high speed and attempts to take a left turn at a road intersection. While turning, a parked vehicle is in the blind spot and left undetected by the autonomous car sensors. In such situations, existing driver intention prediction (DIP) methods [18, 37] may fail to recognize the parked vehicle, increasing the risk of a near-miss or collision. Hence, interpretability in DIP models becomes crucial in the aforementioned cases. Interpretable DIP models can reveal why the system overlooked such situations, which can help diagnose failures and improve model learning. An interpretable model can provide high-level explanations that enhance decision-making, fostering greater trust and confidence in autonomous driving technology. Trust is not solely about performance but also the ability to scrutinize, explain, and refine the model's decisions over time, ultimately ensuring safer and more reliable deployment.

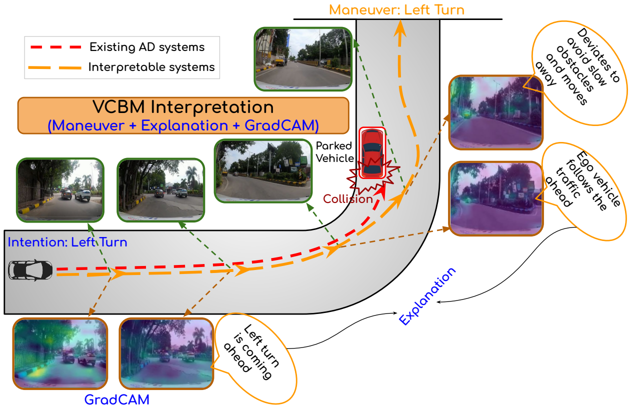
Maneuver: Left Turn
Existing AD systems
Interpretable systems
VCBM Interpretation
(Maneuver + Explanation + GradCAM)
Parked Vehicle
Collision
Intention: Left Turn
Deviates to avoid slow obstacles and moves away
Ego vehicle follows the traffic ahead
Explanation
Left turn is coming ahead
GradCAM

Figure 1. Illustration of an AD scenario for the DIP task. An AD system may intend to take a left turn while encountering a parked or slow-moving vehicle at the turn. Existing DIP models, lacking HCI understanding, might fail to anticipate the obstacle, leading to a potential collision. The proposed interpretable model, VCBM, enhances safety by enabling the ego-vehicle to explain its intended actions, anticipate obstacles more effectively, and adjust maneuvers accordingly. This results in safer and more transparent decision-making.

Traditional DIP datasets such as Brain4Cars [10], Viena^2 [2], HDD [23], AIDE [39], and DAAD [37] primarily focus on predicting maneuvers or agent trajectories without providing contextual explanations. This limitation affects the ability to train and evaluate DIP models on not just *what* happened, but also *why* it happened. To address this gap, we introduce the

DAAD-X dataset (see Table 1), which includes both driving maneuvers (*what*) and corresponding explanations (*why*), enabling richer interpretability.

However, due to their architectural limitations, existing DIP architectures cannot be directly employed to leverage such explanations effectively. For instance, recent architectures, such as VideoMAE [32], DINOv2 [21], and MViTv2 [17] encode spatial and temporal information as flattened token representations, making it challenging to extract human-interpretable insights. Although self-supervised tasks such as frame ordering and motion prediction help capture temporal dynamics, the learned features often fail to correspond intuitively to human-understandable concepts. These limitations highlight the need for models that explicitly align learned features with explanations, ensuring both maneuver prediction and interpretability are jointly optimized.

We address this problem by including concept bottleneck models (CBM) [14], which are widely used to make models interpretable. CBM converts the highly uninterpretable features to low-dimensional, human-understandable explanations by training every neuron of a layer to represent one explanation. These explanations are fed to a sparse linear layer for the final model prediction. This results in a more straightforward interpretation of the final model prediction through linear combinations of interpretable explanations. However, applying these CBMs to video tasks is non-trivial since it does not understand the inherent temporal context of video data, a gap largely unexplored in the literature.

To overcome these challenges, we propose video CBM. By integrating spatially and temporally consistent tokens with CBMs, our approach delivers high-level explanations that naturally capture spatio-temporal features, offering the best of both worlds. To improve the understanding of DIP models, we make the following contributions in this work:

- We propose DAAD-X, a multi-modal driving action anticipation video dataset incorporating hierarchical in-cabin eye-gaze and out-cabin ego-vehicle explanations. This dataset provides human-understandable justifications for driving maneuvers, enhancing interpretability and decision-making transparency.
- Propose a multi-modal video-aware concept bottleneck model (VCBM) with learnable token merging and localized concept bottleneck. Our approach effectively leverages spatio-temporal features to disentangle explanations. To the best of our knowledge, this is the first work to propose a concept-based interpretability method explicitly tailored for video-based models.
- We presented qualitative results for VCBM on the DAAD-X dataset, demonstrating its improved performance across multiple backbone models. In addition, we introduced a multi-label t-SNE visualization to highlight the causal correlation between multiple explanations in a video, offering a deeper interpretation of the model's reasoning.

Table 1. Comparison of datasets. Our dataset is a subset of DAAD [37] and it includes additional categories of explanations for multi-modal videos, encompassing both in-cabin (Aria eye-gaze) and out-cabin (ego-vehicle) perspectives.

Dataset	**#In-cabin view**	**#Out-cabin view**	**Multimodal data**	**Video data**	**Eye-gaze**	**eX-Annotation**	**eX-Temporal /Frame**	**eX-Scene**	**eX-PoV**	**eX-Semantic**	**eX-Causality**
HDD [23]	0	1	No	Yes	✗	N/A	N/A	✗	✗	✗	✗
ROAD [29]	0	1	No	Yes	✗	N/A	N/A	✗	✗	✗	✗
Dr(eye) [22]	1	1	Yes	Yes	✓	N/A	N/A	✗	✗	✗	✗
DAAD [37]	2	4	Yes	Yes	✓	N/A	N/A	✗	✗	✗	✗
BDD-OIA [38]	0	1	No	No	✗	Categorical	Frame	✓	✗	✗	✗
BDD-X [13]	0	1	No	Yes	✗	Contextual	Temporal	✓	✗	✓	✓
DAAD-X (Ours)	2	4	Yes	Yes	✓	Categorical	Temporal	✓	✓	✓	✓

*eX means explanation

2. Related Work

2.1. Driver Intention Prediction

Various methods have been explored to recognize ego-vehicle actions and driver intentions. Early approaches, such as Hidden Markov Models [33], focused on vehicle state prediction, while recent research has shifted to deep learning-based driver action anticipation. Traditionally, bidirectional RNNs [20] and CNN-LSTM architectures [9], [10], [11], [12], [25], [3] were used, though they often emphasize spatial features over temporal dependencies, limiting performance in extended video sequences. To address this, transformer-based architectures [34] were introduced, improving long-range dependency capture, and memory-based anticipation methods such as Cemformer [18] and M^{2}MVIT [37] enhanced temporal consistency. One closely related work [38] produces explanations only for a single frame without incorporating the temporal context. The explanations are limited to short words or phrases, lacking the granularity required to capture cross-frame dynamics, making them unsuitable for interpreting video models.

These video models remain uninterpretable, posing challenges for safe deployment in AD systems. To address this limitation, we propose a video-based interpretable intention prediction model with human-understandable explanations.

2.2. Explainable Video Dataset

Interpretability has recently gained significant attention, yet video-based interpretability remains challenging in tasks such as action recognition and long-video understanding [8, 36]. Existing DIP datasets, such as Brain4Cars [10], Viena^2 [2], HDD [23], AIDE [39], and DAAD [37], though they provide maneuver labels under diverse conditions, lack reasoning or explanatory annotations, limiting their suitability for interpretable models. BDD-OIA [38] offers single frame-level explanations but fails to capture the spatio-temporal context necessary for comprehensive intention prediction models across a video. Although BDD-X [13] offers detailed freeform contextual explanations, it cannot create interpretable models since we require categorical annotations to link a particular driving action to a specific, repeatable explanation with precise, distinct mapping. To bridge this gap, we introduce a new multi-modal video-based driving action dataset with human-understandable explanations to advance interpretability in autonomous driving.

2.3. Concept-Based Explanations

Understanding model decisions in multi-modal and temporal contexts is challenging due to the added temporal dimension and complex shared representations [24]. Prior works, including concept bottleneck models [14, 19], and concept relevance propagation [1], use fixed human-understandable concepts for decision-making, offering interpretability but failing to model temporal inputs [15]. This limitation can lead to models learning spurious correlations and overlooking non-linear feature relationships. Due to the rigorous requirements of manual human-understandable annotations for CBM, the label-free CBMs were introduced in [19, 27, 31] to generate concepts with the help of a pretrained text encoder. Recent methods such as LaIAR [36] and HENASY [35] have used language grounding on videos for contextual interpretability, but they fail in driving tasks due to the inability of the language model to capture positional and directional cues, which are essential in driving. To address this issue, we propose a simple framework that pools relevant features across frames, generating fine-grained, faithful explanations for videos.

3. DAAD-X Dataset

Motivation: While driving on a straight, smooth road, a driver typically makes minimal steering adjustments or eye movements, as fewer decisions are required. However, during maneuvers—such as turning, lane changing, or stopping—the driver must be highly attentive, making precise hand movements based on visual cues. In such critical moments, DIP models can predict actions (e.g., turning left, slowing down), but they do not inherently explain why a particular action was predicted or whether it was the correct decision. For instance, consider a scenario where a driver approaches an intersection. If the DIP model predicts a left turn but does not indicate whether the decision was influenced by a traffic signal, movement of another vehicle , or presence of pedestrians, the prediction remains a black box. Without explanations, assessing whether the model's reasoning aligns with human decision-making is difficult.

To bridge this gap, it is essential to annotate DIP datasets with both actions and corresponding explanations, i.e., both ego-vehicle and eye-gaze explanations. By incorporating these explanations—such as the presence of obstacles, road signs, or the driver's eye-gaze behavior—we can develop interpretable models that not only predict actions and intentions but also justify their decisions. This would enhance trust, safety, and usability of autonomous driving.

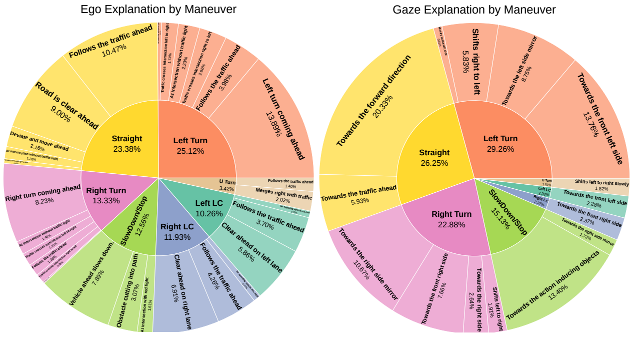

Figure 2. Driving video annotation statistics of DAAD-X dataset. Illustrating the distribution of (left) ego-vehicle explanations and (right) eye-gaze explanations across different maneuver actions. Details of the full explanation annotation are provided in the *Supplementary Material*. Zoom in for better clarity.

Dataset details: To address these issues, we created a new dataset from the DAAD dataset [37] to generate human-understandable explanations. DAAD was the closest match for our setup, as it is multi-modal with eye-gaze information and is well-conditioned across diverse weather conditions, drivers, times of day, and driving scenarios. The DAAD dataset includes seven intention labels, each corresponding to a specific maneuver: go straight (ST), right turn (RT), left turn (LT), right lane change (RLC), left lane change (LLC), slow/stop (SS), and U-turn (UT). We selected 1,568 video clips from the DAAD dataset, each ranging from 7 to 15 seconds in length. These videos are annotated with 17 ego-vehicle explanations and 15 gaze explanations using the VIA video annotator [6], an open-source tool. We refer to this enriched dataset as the explainable DAAD dataset (DAAD-X). Table 1 compares DAAD-X with previous datasets.

3.1. Data Annotation and Statistics

Annotation details: During annotation, annotators watch each driving video and assign reasoning for the driver's maneuver. This process involves selecting a relevant gaze explanation and one or more ego-vehicle explanations to provide contextual justification. The gaze explanation is a single-attribute label chosen from 15 predefined gaze explanations, which indicate where the driver is looking based on gaze coordinates collected using the Aria eye tracker [30]. In contrast, ego-vehicle explanations consist of 17 multi-attribute labels, where multiple explanations can be assigned to a single video. These explanations capture key scene attributes and offer semantically meaningful cues for both spatial and temporal localization. For example, in the explanations "ego-vehicle is nearing the intersection", "road is clear ahead on the left lane", "gaze is mostly towards front right side", the term "nearing" conveys temporal semantics, while "clear ahead" and "front right side" provide spatial context from the ego-vehicle's perspective.

The annotated DAAD-X dataset provides a rich set of explanations, capturing both gaze and ego-vehicle attributes to

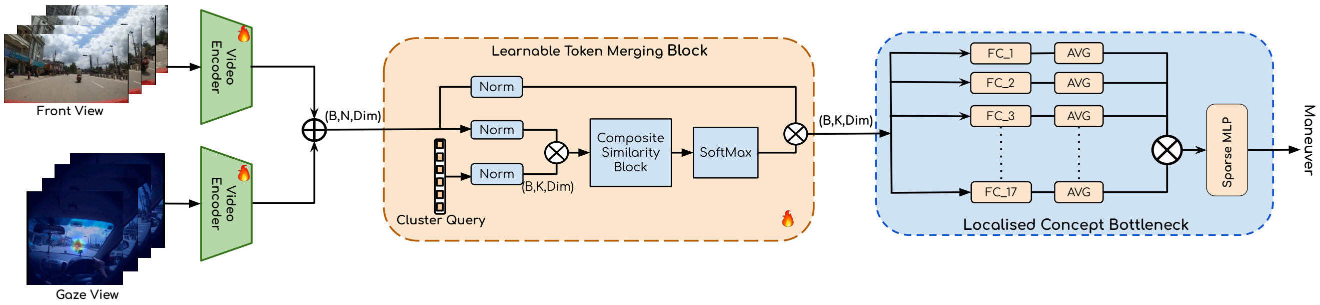

Figure 3. Overall architecture of the proposed VCBM. The dual video encoder first generates the spatio-temporal features (tubelet embeddings) for the ego-vehicle and gaze input sequence video pair. These tubelets are concatenated along the channel dimension and fed into the proposed learnable token merging block to produce K-cluster centers based on composite distances. These clusters are then fed into a localised concept bottleneck to disentangle and predict the maneuver label and one or more explanations to justify the maneuver decision.

enhance interpretability. As illustrated in Fig. 2, each annotated instance includes the driver's maneuver, a gaze explanation, and multiple ego-vehicle explanations. In total, the dataset contains 2,536 explanations, though their distribution is highly unbalanced. Among gaze explanations, the most frequent is "towards the forward direction" (223 occurrences), while the least common is "to the left side" (10 occurrences). Similarly, for ego-vehicle explanations, "a left turn coming ahead" appears most frequently (199 occurrences), whereas explanations such as "nearing an intersection and traffic light is green" are rare, occurring only 7 times for the go straight maneuver, 6 times for a left turn, and just 2 time for a right turn. Given this long-tail distribution, we apply stratified sampling to ensure balanced representation, splitting the dataset into training (70%), validation (20%), and testing (10%) sets.

Sanity Check. Once annotated, the annotations were shuffled 3 times among the annotators to validate the explanations. Since explanations are subjective, we initially selected the most obvious ones. For ambiguous cases, a consensus was reached based on the votes and comments of 10 annotators. With this process, less than 1% of the total videos were found to be incorrectly annotated and subsequently corrected. More details in *Supplementary Material*.

4. Video Concept Bottleneck Model (VCBM)

4.1. Problem Formulation

Given a set of input videos, where each video consists of $x_g, x_f \in R^d$. x_g represents the gaze view video and x_f is a front view of an ego-vehicle. Now, we have a corresponding driving maneuvers prediction for each video sequence represented by y and explanations denoted as e. $e \in \{0,1\}^{17}$ represents the 17 explanations. Consider the training dataset consisting of $\{(x_g^i, x_f^i, y^i, e^i)\}_{i=1}^{T}$, where T is the total training instances. We can predict $y = f(g(x))$ where $g : R^d \to R^{17}$ represents the bottleneck layer, which maps input video features to 17 intermediate explanations. $f : R^{17} \to R$ is a sparse linear layer that maps intermediate explanations to the final maneuver prediction labels.

In our work, we introduce an unsupervised clustering module $m : R^d \to R^d$ (detailed discussion in Section 4.3) to cluster similar features across frames. Finally, we follow [14] to learn bottleneck model $(\hat{f}, \hat{m}, \hat{g})$ using the joint bottleneck approach, which minimizes the weighted sum as,

$$\hat{f},\hat{m},\hat{g} = \text{argmin}_{f,m,g}\left(\sum_i \left[L_Y\left(f(g(m(x^{(i)}))),y^{(i)}\right) + \sum_j \lambda L_{C_j}\left(g(m(x^{(i)})),e_j^{(i)}\right)\right]\right) \tag{1}$$

L_y and L_{C_j} represent multiclass cross entropy loss and multilabel aggregated binary cross entropy loss for each explanation $j \in \{1,17\}$. Here, λ is the weighting factor.

4.2. Our Model Architecture

We illustrate VCBM, our proposed model architecture, in Fig. 3. VCBM comprises a dual video encoder, a novel learnable token merging (LTM), and a localised concept bottleneck model (LCBM) module. LTM and LCBM help interpret both in-cabin gaze and out-cabin front video data effectively. At its core, VCBM predicts the driver's intended maneuver and provides human-understandable explanations for why the maneuver was selected, enhancing interpretability in the DIP task.

Video Encoder. Our video encoder architecture is based on [34]. For an input video sequence, gaze video as x_g^i and ego-vehicle front video as x_f^i, we pass them individually to each branch to extract individual feature embeddings $z_i = (z_g, z_f) \in R^{B \times N \times Dim}$. B, N, Dim represents the batch size, number of tokens, and feature representation dimension. To retain spatial positioning while maintaining temporal consistency [40], we concatenate these feature embeddings along the channel dimension represented as z_i'. Now, we pass z_i' through our proposed modules, Learnable Token Merging (LTM) and Localised Context Bottleneck Model (LCBM), to obtain the final prediction discussed in detail in the remaining section.

4.3. Learnable Token Merging

We introduce an LTM module to ensure the LCBM captures local features across frames. LTM groups semantically similar features into a reduced set of representative tokens, which are then passed as inputs to LCBM. We first perform unsupervised clustering in LTM on the concatenated multi-view feature representations $z_i^{'}$ from the encoder. The features are compared with K learnable cluster centers z_{c_j}, where (i,j) represents token position and $K << N$ to ensure that explanations are assigned to a compact set of merged features. The similarity between a feature token and a cluster center is computed using cosine similarity, given by,

$$d_{feat}^{(i,j)} = 1 - \frac{z_i' \cdot z_{c_j}}{\|z_i'\| \|z_{c_j}\|}, \quad \forall i \in \{1,...,N\}, \quad \forall j \in \{1,...,K\} \tag{2}$$

We introduce a composite similarity block (in Fig. 3) that integrates and refines the similarity measures. The composite similarity block enhances clustering by enforcing spatial and temporal consistency. We compute additional spatial $\tilde{d}_{spatial}^{(i,j)}$ and temporal distances $\tilde{d}_{temporal}^{(i,j)}$,

$$\tilde{d}_{spatial}^{(i,j)} = \frac{d_{spatial}^{(i,j)}}{S_{\max}}, \quad d_{spatial}^{(i,j)} = \sqrt{(x_i - x_{c_j})^2 + (y_i - y_{c_j})^2} \tag{3}$$

$$\tilde{d}_{temporal}^{(i,j)} = \frac{d_{temporal}^{(i,j)}}{T_{\max}}, \quad d_{temporal}^{(i,j)} = |t_i - t_{c_j}| \tag{4}$$

The total composite distance used for clustering is,

$$d_{composite}^{(i,j)} = \alpha d_{feat}^{(i,j)} + \beta \tilde{d}_{spatial}^{(i,j)} + \gamma \tilde{d}_{temporal}^{(i,j)} \tag{5}$$

Here, the x_c, y_c, t_c represent the learnable cluster center position in spatial and temporal dimensions, respectively, and α, β, γ represent the normalized weights for distances. Instead of hard clustering [7, 28], we employ soft clustering by assigning soft labels w_{ij} to each token z_i using a softmax over the negative composite distances,

$$w_{ij} = \frac{exp(-d_{composite}^{(i,j)})}{\sum_{j=1}^{K} exp(-d_{composite}^{(i,j)})} \tag{6}$$

The updated cluster centers are then computed as a weighted sum of token embeddings,

$$\tilde{z}_{c_j} = \frac{\sum_{i=1}^{N} w_{ij} z_i}{\sum_{i=1}^{N} w_{ij}} \tag{7}$$

By merging similar features into a compact token representation, this approach reduces redundancy in video embeddings while ensuring that spatio-temporally relevant features are retained. These merged token representations serve as inputs to the LCBM, enabling it to generate fine-grained explanations while maintaining spatial and temporal consistency.

4.4. Localised Context Bottleneck Model

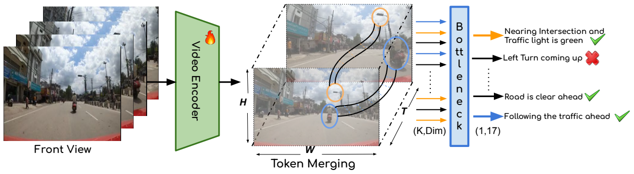

Figure 4. VCBM merges relevant features across frames (z_{c_j}) and assigns explanations. Blue represents **merged traffic features**, **orange** denotes **merged traffic light features**, and arrow thickness indicates prediction confidence.

The LCBM further refines the representations from LTM by mapping high-dimensional encoded vectors to a human-understandable low-dimensional space. Traditional CBM approaches rely on global feature embeddings or global average pooling, which can discard fine-grained spatial and temporal details. Instead, LCBM approach preserves these details by feeding all the pooled token representations into the bottleneck block ($g(z_c)$), as illustrated in Fig. 4.

Rather than immediately averaging features before the bottleneck, we introduce a late averaging strategy, allowing each merged token to retain its individual contribution to the explanation process. Each fully connected (FC) layer in the bottleneck module corresponds to a specific explanation and produces a single logit, representing the confidence of that explanation. This ensures that each FC layer processes all tokens, enabling a more robust and interpretable assignment of explanations. By retaining fine-grained spatio-temporal details, this LCBM enhances the activation maps, leading to more precise and human-understandable explanations in DIP.

5. Experiments

5.1. Implementation Details

For our experiments, we used I3D [4] pre-trained on ImageNet RGB images, as well as VideoMAE [32] with a ViT-B/16 backbone [5] and MViTv2-B [17], both pre-trained on Kinetics-400 dataset. For more details on data augmentation, training parameters, and evaluation metrics refer *Supplementary Material*.

5.2. Results

We compare our proposed method with three backbone architectures: the CNN-based I3D, transformer-based VideoMAE, and MViTv2. Table 2 shows the performance of the baselines and the backbone models with and without the bottleneck layer. We observe that the transformer-based MViTv2 baseline outperforms the CNN-based I3D baseline in predicting explanations using the bottleneck layer. While CNN excels in spatial feature extraction, video-based explanation tasks require a strong temporal understanding across frames, making transformers more effective.

Table 2. Evaluation on DAAD-X dataset: Evaluated baselines with (wB) and without (woB) bottleneck. Here, LTM indicates Learnable Token Merging.

Model	Action		ego-vehicle eXplanation			
	Acc	F_1	Acc	F_1	$F_1(mac)$	$F_1(mic)$
I3D woB [4]	74.78	74.21	-	-	-	-
VideoMAE woB [34]	72.5	71.81	-	-	-	-
MViTv2 woB [17]	64.03	63.98	-	-	-	-
I3D wB [4]	74.09	73.47	25.26	36.73	18.53	43.49
VideoMAE wB [34]	67.01	66.48	24.21	38.24	23.77	41.53
MViTv2 wB [17]	63.29	62.47	25.35	37.1	24.3	42.1
I3D + LTM wB (Ours)	**73.21**	**72.2**	**28.31**	**39.43**	**24.1**	**44.06**
MViTV2 + LTM wB (Ours)	**69.73**	**69.15**	**31.22**	**43.86**	**29.17**	**49.11**

Table 3. Importance of token aggregation. Compared the effectiveness of token merging (using all the tokens from the encoder) over utilizing the CLS token for producing explanations.

Strategy	Model	Action		ego-vehicle eXplanation			
		Acc	F_1	Acc	F_1	$F_1(mac)$	$F_1(mic)$
CLS Token Summarization	I3D woB	74.38	74.1	-	-	-	-
	I3D wB	73.73	72.25	23.15	33.7	16.9	41.95
Full Token Aggregation*	I3D woB	**74.78**	**74.21**	-	-	-	-
	I3D woB	73.21	72.2	**28.31**	**39.43**	**24.1**	**44.06**

Table 4. Effect of number of clusters. A bottleneck with lower clusters learns more global representation. Increasing the clusters further reduces performance due to noisy cluster centers.

Model	#Clusters	Action		ego-vehicle eXplanation			
		Acc	F_1	Acc	F_1	F_1 (mac)	F_1 (mic)
I3D + LTM wB	1	70.78	70.47	24.64	34.47	16.67	42.29
I3D + LTM wB	3	71.78	71.44	25	38.46	22.85	43.87
I3D + LTM wB	5	73.21	72.2	**28.31**	**39.43**	**24.1**	**44.06**
I3D + LTM wB	7	**74.28**	**74.03**	24.64	34.9	18.44	41.01
I3D + LTM wB	10	70.35	69.64	23.92	33.28	16.09	41.2
MViTV2 + LTM wB	5	**69.73**	**69.15**	**31.22**	**43.86**	**29.17**	**49.11**
MViTV2 + LTM wB	10	65	64.53	30	43.51	27.11	47.11

5.3. Insights and Ablations

In this section, we provide additional insights into the effect of token merging, the importance of gaze modality in VCBM, and the impact of temporal cues.

5.3.1. Effects of LTM and LCBM

In Table 4, we analyze the impact of varying the number of clusters in the LTM block (see Fig. 3) on both action and explanation predictions in the proposed method. Using a single cluster is analogous to employing a CLS token in a transformer, where all tokens are aggregated into one global representation. As the number of clusters increases, each cluster is attributed to certain similar features across the tokens, but adding more clusters counteracts by learning additional noise patterns, which detracts from the prediction performance.

Table 5. Component-level ablation. Significance of proposed modules (LTM and LCBM) on I3D architecture.

Components		Action		ego-vehicle eXplanation			
LTM	LCBM	Acc	F_1	Acc	F_1	$F_1(mac)$	$F_1(mic)$
✗	✗	68.1	67.44	11.22	21.44	9.37	22.51
✓	✗	72.8	72.15	26.03	35.6	19.15	44.1
✗	✓	**74.09**	**73.47**	25.26	36.73	18.53	43.49
✓	✓	73.21	72.2	**28.31**	**39.43**	**24.1**	**44.06**

LCBM is designed to compute explanations based on all locally aggregated tokens from the LTM block, rather than relying on a single global CLS token. This approach enhances the justification of explanations by preserving the contextual integrity of feature groups, making the model more interpretable as illustrated in the Table 3. The component-level ablation shown in Table 5 demonstrates that incorporating the LCBM block improves explanation and action performance metrics. By attending to all the input tokens, the LCBM effectively retains fine-grained details, leading to more precise bottleneck representations. Additionally, integrating the LTM block further enhances explanation performance due to its ability to extract meaningful merged tokens. However, the averaging process during token merging reduces the granularity of individual features, resulting in a slight drop in action prediction accuracy.

5.3.2. Importance of Gaze modality

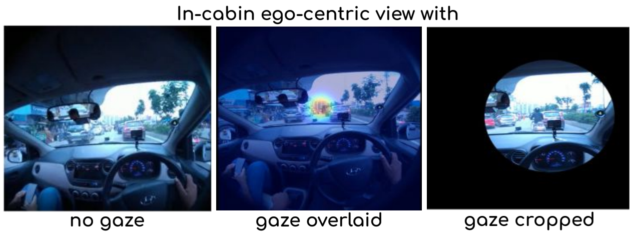

Figure 5. Variants of gaze input. The driver's view video is processed in the following way to show the best way to represent gaze without affecting spatial features. The gaze cropped variant ($R = 350$) produces the best quantitative results.

We analyse the impact of gaze modality on explanation predictions by testing three settings: without gaze, gaze overlaid, and with gaze cropped regions, as illustrated in Fig. 5. In Table 7, we show the accuracy of identifying the gaze variant. Initially, without gaze, both action and explanation predictions are lower. Further, the gaze is being overlaid onto the driver's view, but this method adds noise to the image and deteriorates finer details. To mitigate this issue, we cropped a circular region centered on the gaze ground truth from the driver's view, experiments with various diameters (in pixels), in Table 7 show

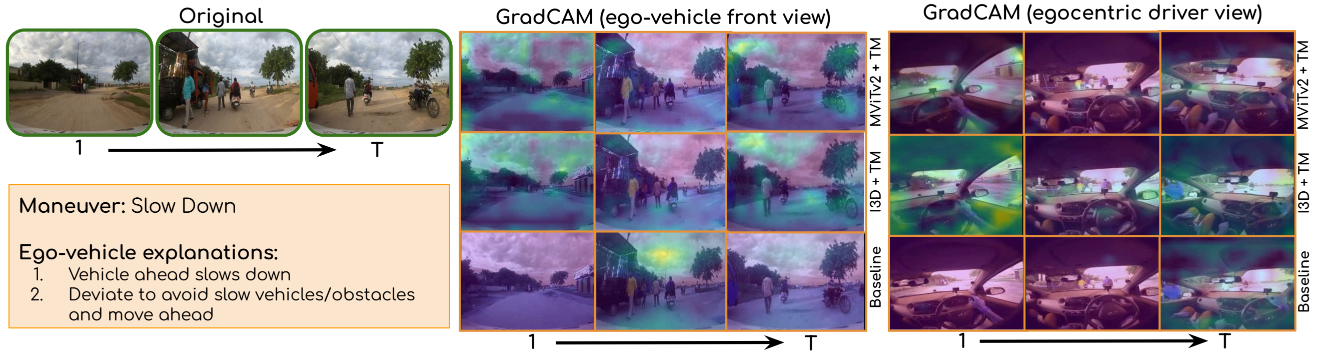
Original
1
T
Maneuver: Slow Down
Ego-vehicle explanations:
1. Vehicle ahead slows down
2. Deviate to avoid slow vehicles/obstacles and move ahead
GradCAM (ego-vehicle front view)
MViTv2 + TM
I3D + TM
Baseline
1
T
GradCAM (egocentric driver view)
MViTv2 + TM
I3D + TM
Baseline
1
T

Figure 6. GradCAM visualization on proposed method. At $t = 1$, the activations are scattered, but as time progresses to $t = T$, the CAM gradually refines and localise on important objects. This represents how humans make decisions, which evolves over time.

improved performance for explanation by achieving an optimal score at $R = 350$. However, excessively increasing the crop size reduced the concentration of relevant gaze information, reducing the explanation performance.

5.3.3. Tradeoff between Explanation and Action Classification

We train explanation classification jointly with action prediction to align with human reasoning. Table 6 shows that adding the auxiliary explanation loss through the scaling parameter λ in Equation 1 boosts both explanation and action accuracy. However, excessive weight can slightly affect the action classification performance.

Table 6. Tradeoff of Explanation Classification. Increasing the emphasis on explanation classification leads to a decline in the action prediction performance.

scaling factor (λ)	Action		ego-vehicle eXplanation			
	Acc	F_1	Acc	F_1	$F_1(mac)$	$F_1(mic)$
0	72.14	71.82	0	14.56	8	8.65
0.01	71.07	69.94	0.71	9.87	3.3	12.95
0.1	**74.28**	**73.43**	5.71	17.36	10.09	14.48
0.5	73.21	72.2	28.31	39.43	**24.1**	44.06
1	70.35	69.23	**30**	**40.02**	22.58	**46.4**

5.3.4. Effect of Temporal Cues

In Fig. 7, we show the impact of temporal cues on action and explanation performance across CNN and transformer models. Interestingly, transformers exhibit lower action prediction accuracy than CNNs, likely due to two factors: (1) CNNs rely more on spatial features and process limited temporal context, suggesting that DIP tasks can be addressed at the frame level but at the cost of explainability, and (2) transformers undergo stronger regularisation to prevent learning spurious correlations from noisy data, with the help of random shuffling as discussed in Sec. 4. This random shuffling disrupts temporal order, affecting transformer performance, while CNNs remain unaffected since they are less reliant on temporal information. The severity s in Fig. 7 represents the degree of reshuffling.

Let T be the total number of frames in the video. We first divide it into 16 equal segments, each containing $\ell = \frac{T}{16}$ frames. Based on the severity parameter s, we merge every s consecutive segments, forming $M = \frac{16}{s}$ merged segments. From each merged segment, we uniformly sample s frames:

$$F_i = \{f_{i,1}, f_{i,2}, \ldots, f_{i,s}\}, \quad i = 1, \ldots, M$$

where each sampled frame $f_{i,j}$ is selected uniformly from the i^{th} merged segment. The total number of sampled frames sums up to:

$$\sum_{i=1}^{M} |F_i| = 16.$$

This indicates that our explanation annotations are influenced by temporal dependencies.

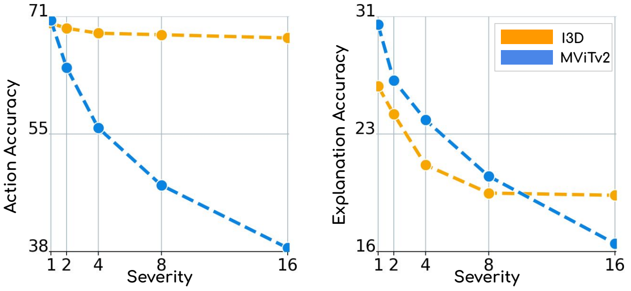

Figure 7. Effect of temporal cues. As the severity of frame reshuffling increases, the action and explanation accuracy of MViTv2 drops significantly compared to I3D. Notably, explanation accuracy drops more than action accuracy, indicating the importance of temporal cues for producing meaningful explanations.

5.4. Qualitative Analysis

5.4.1. GradCAM Visualization

As shown in Fig. 6, the vehicle's current maneuver is predicted as "Slow Down", with intermediate outputs providing ego-vehicle explanations. This demonstrates that LTM and LCBM effectively focus activations on relevant features that align

Table 7. Gaze modality input variants. Having the gaze cropped regions is better than the usual way of overlaying the gaze in the DIP task.

Variants	Action		ego-vehicle eXplanation			
	Acc	F_1	Acc	F_1	$F_1(mac)$	$F_1(mic)$
no gaze	68.11	67.94	8.77	17.52	9.37	22.51
overlaid	71.57	70.57	14.03	22.6	11.86	28.55
50	67.85	67.51	14.28	24.24	11.34	28.8
150	70.21	69.2	20.63	28.46	15.22	34.18
250	72.63	72.42	23.74	33.36	17.59	40.13
350	74.09	73.47	**26.42**	**36.73**	18.53	**43.49**
450	**74.64**	73.74	25.26	36.01	**19.18**	43.46
550	74.28	**73.75**	24.64	35.67	17.39	43.32

with the predicted explanations. In contrast, the baseline CBM, without late averaging and LTM, results in more dispersed, global activations, making it less interpretable.

5.4.2. Label-anchored Multi-label t-SNE Visualization

Visualizing explanations in the feature space is essential for understanding what the DIP model has learned in maneuver prediction. When dealing with multi-label explanations, techniques like t-SNE help interpret the latent space and reveal relationships between different explanations. However, since t-SNE is not directly suited for multi-label classification, we introduce explanations as anchor points within the latent space. This approach offers two key benefits: (1) anchor points highlight the degree of correlation among different explanations, and (2) individual video features are positioned in alignment with and in close proximity to their relevant anchor points, ensuring a more interpretable representation of the learned features.

To formalize, let $\mathbf{z'}_i \in \mathbb{R}^d$ be the backbone feature vector for the i-th sample, for $i = 1,...,T$. For each explanation k, define a mask indicator $s_i^{(k)}$ as follows,

$$s_i^{(k)} = \begin{cases} 1, & \text{if class } k \text{ is activated for sample } i, \\ 0, & \text{otherwise.} \end{cases}$$

The aggregated feature representation for the explanation k, denoted by $\bar{\mathbf{z}}_k$, is calculated by applying the mask to the features and then averaging over the samples where the mask is active,

$$\bar{\mathbf{z}}_k = \frac{\sum_{i=1}^{T} s_i^{(k)} \mathbf{z'}_i}{\sum_{i=1}^{T} s_i^{(k)}},$$

This aggregated feature $\bar{\mathbf{z}}_k$ serves as an anchor in the 2D t-SNE space for explanation k. This makes the feature space more interpretable, revealing degree of causal relationships learned between explanations.

Fig. 8 illustrates the 17 explanation anchors, each marked with distinct shapes. This visualization highlights that semantically related explanations tend to form clusters, while individual video features (depicted as colored points) are positioned near their corresponding explanation anchors. For example, if a video contains both “traffic light is green” and “left turn coming ahead,” its feature representation will be located near both respective anchors, reflecting the model’s learned associations.

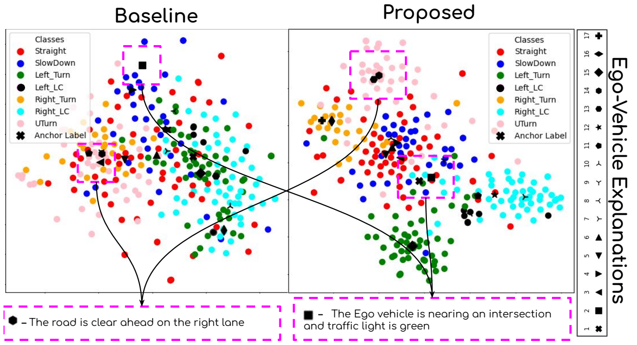

Figure 8. Label-anchored multi-label t-SNE. Colored dots represent clusters of individual video features. **Left**: Baseline model exhibiting a poorly disentangled representation space. **Right**: Proposed method demonstrating improved separation of explanation symbols, with a stronger causal correlation to the videos. The square marker is positioned at the center, representing a feature commonly observed across all videos. The hexagon indicates an explanation learned in scenarios where a U-turn is performed, as a right turn and a right lane change always accompany it.

Limitations and future scope: Our proposed method generates high-level explanations while preserving faithful feature attributes. However, GradCAM activations are predominantly observed in the forward direction. This occurs because we assume that important objects are only considered if they are visible from both views, i.e, the driver’s gaze should guide, where the model should focus on from the front view. Consequently, token merging relies on the assumption that the video frames are at least partially aligned. An interesting direction for future work would be to investigate the effects of explicitly aligning both views before performing token merging using techniques like homography.

6. Conclusion

This work introduces a novel paradigm for conceptual interpretability in driving maneuver prediction. We developed a comprehensive multi-modal dataset incorporating human-understandable explanations to help create interpretable models within the autonomous driving systems. Our analysis of existing architectures revealed that transformer models excel at generating explanations due to their inherent temporal bias. Leveraging this insight, we proposed VCBM, a model that merges spatio-temporal features to predict localised explanations and reliably represents them through post hoc feature attribution methods. Additionally, our feature-level visualization approach effectively elucidates the causal correlations among explanations, enhancing driver intention prediction systems’ overall transparency and reliability.

Acknowledgment. The project was supported by iHub-Data and Mobility at IIIT Hyderabad.

References

[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019, 2023. 3

[2] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. VIENA2: A driving anticipation dataset. *CoRR*, abs/1810.09044, 2018. 1, 2

[3] Mahdi Bonyani, Mina Rahmanian, Simindokht Jahangard, and Mahdi Rezaei. Dipnet: Driver intention prediction for a safe takeover transition in autonomous vehicles. *IET Intelligent Transport Systems*, 2023. 2

[4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pages 4724–4733. IEEE Computer Society, 2017. 5, 6

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021. 5

[6] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *ACM Multimedia*, pages 2276–2279, 2019. 3

[7] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226–231. AAAI Press, 1996. 5

[8] Jorge Garcia-Torres Fernandez. Interpretability in video-based human action recognition: Saliency maps and gradcam in 3d convolutional neural networks. *Nordic Machine Intelligence*, 2024. 2

[9] Patrick Gebert, Alina Roitberg, Monica Haurilet, and Rainer Stiefelhagen. End-to-end prediction of driver intention using 3d convolutional neural networks. *IEEE Intelligent Vehicles Symposium (IV)*, pages 969–974, 2019. 2

[10] Ashesh Jain, Hema S Koppula, Bharad Raghavan, Shane Soh, and Ashutosh Saxena. Car that knows before you do: Anticipating maneuvers via learning temporal driving models. In *ICCV*, pages 3182–3190, 2015. 1, 2

[11] Ashesh Jain, Avi Singh, Hema S Koppula, Shane Soh, and Ashutosh Saxena. Recurrent neural networks for driver activity anticipation via sensory-fusion architecture. In *ICRA*, pages 3118–3125. IEEE, 2016. 2

[12] Nima Khairdoost, Mohsen Shirpour, Michael A Bauer, and Steven S Beauchemin. Real-time driver maneuver prediction using lstm. *IEEE Transactions on Intelligent Vehicles*, 5(4): 714–724, 2020. 2

[13] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. *ECCV*, 2018. 2, 3

[14] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, pages 5338–5348, 2020. 2, 3, 4

[15] Jae Hee Lee, Georgii Mikriukov, Gesina Schwalbe, Stefan Wermter, and Diedrich Wolter. Concept-based explanations in computer vision: Where are we and where could we go? *CoRR*, abs/2409.13456, 2024. 3

[16] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *CVPR*, pages 7644–7652, 2019. 1

[17] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, pages 4804–4814, 2022. 2, 5, 6

[18] 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. 1, 2

[19] Tuomas P. Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *ICLR*, 2023. 3

[20] Oluwatobi Olabiyi, Eric Martinson, Vijay Chintalapudi, and Rui Guo. Driver action prediction using deep (bidirectional) recurrent neural network. *CoRR*, abs/1706.02257, 2017. 2

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and et al. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 2

[22] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. Predicting the driver's focus of attention: The dr(eye)ve project. *IEEE TPAMI*, 41(7):1720–1733, 2019. 2

[23] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7699–7707, 2018. 1, 2

[24] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *ACM SIGKDD*, pages 1135–1144, 2016. 3

[25] Yao Rong, Zeynep Akata, and Enkelejda Kasneci. Driver intention anticipation based on in-cabin and driving scene monitoring. In *IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8, 2020. 2

[26] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019. 1

[27] Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. Incremental residual concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11040, 2024. 3

[28] Kristina P. Sinaga and Miin-Shen Yang. Unsupervised k-means clustering algorithm. *IEEE Access*, 8:80716–80727, 2020. 5

[29] Gurkirt Singh, Stephen Akrigg, Manuele Di Maio, Valentina Fontana, Reza Javanmard Alitappeh, Salman Khan, Suman Saha, Kossar Jeddisaravi, Farzad Yousefi, Jacob Culley, and et al. Road: The road event awareness dataset for autonomous driving. *IEEE TPAMI*, 45(1):1036–1054, 2023. 2

[30] Kiran K. Somasundaram, Jing Dong, Huixuan Tang, Julian Straub, Mingfei Yan, Michael Goesele, Jakob Julian Engel, Renzo De Nardi, and Richard A. Newcombe. Project aria: A new tool for egocentric multi-modal AI research. *CoRR*, 2023. 3

[31] Andong Tan, Fengtao Zhou, and Hao Chen. Explain via any concept: Concept bottleneck model with open vocabulary concepts. In *European Conference on Computer Vision*, pages 123–138. Springer, 2024. 3

[32] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2, 5

[33] Duy Tran, Weihua Sheng, Li Liu, and Meiqin Liu. A hidden markov model based driver intention prediction system. In *IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 115–120, 2015. 2

[34] 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. 2, 4, 6

[35] Khoa Vo, Thinh Phan, Kashu Yamazaki, Minh Tran, and Ngan Le. Henasy: Learning to assemble scene-entities for interpretable egocentric video-language model. *Advances in Neural Information Processing Systems*, 37:86483–86499, 2025. 3

[36] Ning Wang, Guangming Zhu, HS Li, Liang Zhang, Syed Afaq Ali Shah, and Mohammed Bennamoun. Language model guided interpretable video action reasoning. In *CVPR*, pages 18878–18887, 2024. 2, 3

[37] Abdul Wasi, Shankar Gangisetty, Shyam Nandan Rai, and C. V. Jawahar. Early anticipation of driving maneuvers. In *ECCV*, 2024. 1, 2, 3

[38] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. *CVPR*, pages 9520–9529, 2020. 1, 2

[39] Dingkang Yang, Shuai Huang, Zhi Xu, Zhenpeng Li, Shunli Wang, Mingcheng Li, Yuzheng Wang, Yang Liu, Kun Yang, Zhaoyu Chen, et al. Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In *ICCV*, pages 20459–20470, 2023. 1, 2

[40] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *CVPR*, pages 8850–8860, 2024. 4