

# IDD-X: A Multi-View Dataset for Ego-relative Important Object Localization and Explanation in Dense and Unstructured Traffic

Chirag Parikh<sup>1</sup>, Rohit Saluja<sup>2</sup>, C.V. Jawahar<sup>1</sup>, Ravi Kiran Sarvadevabhatla<sup>1</sup>

**Abstract**—Intelligent vehicle systems require a deep understanding of the interplay between road conditions, surrounding entities, and the ego vehicle’s driving behavior for safe and efficient navigation. This is particularly critical in developing countries where traffic situations are often dense and unstructured with heterogeneous road occupants. Existing datasets, predominantly geared towards structured and sparse traffic scenarios, fall short of capturing the complexity of driving in such environments. To fill this gap, we present **IDD-X**, a large-scale dual-view driving video dataset. With 697K bounding boxes, 9K important object tracks, and 1-12 objects per video, **IDD-X** offers comprehensive ego-relative annotations for multiple important road objects covering 10 categories and 19 explanation label categories. The dataset also incorporates rearview information to provide a more complete representation of the driving environment. We also introduce custom-designed deep networks aimed at multiple important object localization and per-object explanation prediction. Overall, our dataset and introduced prediction models form the foundation for studying how road conditions and surrounding entities affect driving behavior in complex traffic situations.

## I. INTRODUCTION

Understanding the influence of road and traffic conditions on ego vehicle’s driving behavior is crucial for enabling explainability in automated driving decision-making. This capability helps in developing reliable and efficient intelligent vehicle systems. Unlike Western countries, developing nations contain dense and unstructured traffic situations, with heterogeneous road occupants (two-wheelers, animals, three-wheelers, etc), and static road objects (speed breakers, potholes, and traffic lights, etc.). Given the complex variety of road occupants and objects, it is important to know which of the multiple road entities influence the ego vehicle’s driving behavior and more importantly, how they do so.

Existing datasets capture annotations for identifying important road entities [1], [2], [3], with accompanying explanations in some cases [4], [5], [6], [3]. In general, these datasets contain structured and sparsely populated traffic situations. In most cases, only a single road occupant is responsible for influencing the ego vehicle’s driving behavior. On the contrary, in dense traffic scenarios found in developing countries, it is common to find multiple important road entities simultaneously affecting the ego vehicle’s driving behavior. In structured driving scenarios, most vehicles adhere to traffic laws due to which there is limited variability in the ways they interact with the ego vehicle. However, road occupants usually do not obey the traffic rules in unstructured traffic scenarios. As a result, much more diverse variety of atypical and unexpected interaction patterns tend to exist.

To ensure the representation of such scenarios among the datasets, we contribute a large-scale dual-view driving video dataset, **IDD-X**, captured in dense, heterogeneous, and unstructured traffic environments. The proposed dataset provides ego-relative annotations for multiple important road objects, their corresponding explanation labels, and the ego vehicle’s driving behavior for the duration of the video clip. Drivers routinely utilize rearview information in their assessment of important objects. To account for this important aspect, **IDD-X** additionally captures road objects perceived in the rearview. By encompassing both front and rear views, **IDD-X** enables a more comprehensive analysis of driving behavior, providing a panoramic view of objects, their interactions, and the intricate cues that influence the driver’s choices.

Harnessing the potential of our **IDD-X** dataset, we also introduce novel deep network architectures to address two key tasks - (1) multiple important object localization and (2) per object explanation prediction. These tasks serve as foundational components for unraveling the nuanced relationships between road conditions and ego vehicle’s driving behavior in diverse and challenging traffic contexts. Our contributions are summarised below.

- 1) We contribute **IDD-X**, the first ever driving video dataset with ego-relative annotations for both multiple important object localization and corresponding explanations in dense and unstructured traffic. **IDD-X** contains 697K bounding boxes, 9K important object tracks, and 1-12 objects per video and comprehensive ego-relative annotations for multiple important road objects (10 categories) and explanations (19 categories).
- 2) In **IDD-X**, we provide the first dataset with multi-view important object annotations for ego vehicle’s driving behavior understanding.
- 3) We introduce custom-designed deep networks for a) multiple important object localization and b) per object explanation prediction with respect to the ego vehicle.

Visit our project page <https://idd-x.github.io> for details.

## II. RELATED WORKS

**Important Object Identification:** Identifying important road entities that influence the ego vehicle’s driving behavior forms an important component of the generated behavior explanations. Approaches for important object identification use the ego vehicle’s trajectory, the object’s trajectory, and the video context information [2] [1]. However, these methods do not consider the category of road objects. Given the heterogeneous road object categories in our dataset, we explicitly incorporate object class data along with trajectory information for predicting its importance.

<sup>1</sup>Centre for Visual Information Technology, IIT Hyderabad, India  
chirag.parikh@research.iit.ac.in

<sup>2</sup>Department of Computer Science, IIT Mandi, Mandi, India  
rohit@iitmandi.ac.in

TABLE I: Comparison of IDD-X with public driving behavior understanding datasets. The I.O. abbreviations used in column names refer to the ego-relative Important Object. The column "Irregular Road Surface Description" refers to speed breakers and potholes as important object annotations.

Dataset	Multi-View	I.O. Location	I.O. Track	Irregular Road Surface Description	Dense Traffic	Unstructured Traffic	#I.O. Bounding Boxes	#I.O.s Per Video	#I.O. Explanation Categories	#I.O. Categories	#I.O. Explanation Type
IDD-X	✓	✓	✓ (9K)	✓	✓	✓	697K	1-12	19	10	Categorical
DRAMA [3]	-	✓	-	-	-	-	17K	1	-	3	Textual
H3D [1]	-	✓	-	-	-	-	8K	1-3	-	4	-
BDD-OIA [4]	-	-	-	-	✓	-	-	-	8 <sup>1</sup>	6	Categorical
OIE [2]	-	✓	-	-	-	-	4K	1-2	-	2	-
HDD [6]	-	-	-	-	-	-	-	1	5	4	Categorical
BDD-X [5]	-	-	-	-	-	-	-	-	-	-	Textual
METEOR [7]	-	-	-	-	✓	✓	-	-	-	-	-

**Important Object Explanation:** Explaining the ego vehicle’s driving behavior in the light of identified important objects is crucial for reliable usage of vision-based driving models [8] [9] [10] [11]. However, existing approaches [5] [12] [13] [14] do not identify important objects when generating explanations. Specifically, they do not localize road entities that form the basis for the explanation. This introduces ambiguity in the evaluation of generated explanations. The model proposed by Kim et al. [5] and Jing et al. [13] provide pixel-level saliency maps along with explanations. However, these maps do not overlap reliably with causal road entities. Recently proposed approaches [3] [4] provide object-level explanations using an RGB image or its optical flow image as input. However, these image-based methods do not utilize temporal information or trajectory of important objects for explanation prediction. In our proposed approach, we extract spatiotemporal features of important object tracks and utilize them for generating ego vehicle explanations.

**Datasets:** Multiple datasets exist for identifying risk objects [3], anomalous objects [15], causal objects [6], and important objects [2], [1] that affect the ego vehicle’s driving behavior. However, the traffic scenarios in these datasets are typically sparse and mostly contain a single important object’s annotation in a driving sequence. Such datasets cannot be used to explain the dense traffic conditions where multiple road objects simultaneously influence the ego’s driving decisions. Although some datasets specifically capture dense traffic scenes [4] [7] [16], they do not provide category and location information for ego-relative important objects. In contrast, our IDD-X dataset has spatial and temporal location annotations for *multiple* important road objects that explain the ego’s driving situation in dense traffic.

Our dataset offers explanations for challenging unstructured traffic situations containing complex and unpredictable interaction patterns which are often rare in structured driving scenarios found in existing datasets. Additionally, IDD-X is the first dataset to consider rearview information for important object annotations – see Table I for a detailed comparison.

<sup>1</sup>For a fair comparison with our dataset, explanations referring to at least one important or action-inducing object are considered and similar explanations for left and right turns are merged into one category.

### III. THE DATASET

Our dataset IDD-X is specifically designed to explain ego vehicle’s driving behavior due to multiple important road objects. These objects affect the ego vehicle’s driving decision and it is important to know the identity and role of the respective objects. To this end, we utilize front and rearview driving videos captured from vehicle-mounted cameras to assess the ego’s surrounding traffic situation. Temporal modeling of road objects is crucial for describing complex and unpredictable driving maneuvers commonly observed in unstructured traffic. Therefore, individual object-level tracks are included. Our dataset also provides annotations for 10 different road object categories: car, motorcycle, autorickshaw, truck, bus, bicycle, person/animal, speed breaker and damaged road (pothole). If an object does not fall into any of the listed categories, it is assigned to ‘others’ class.

Our dataset considers the following predominant driving situations: a) slowing down on straight roads, b) deviating on straight roads, and c) slowing down on left/right/U-turns. Every important object is associated with an ego-relative explanation during the driving situation. The explanations in IDD-X comprise a set of pre-defined categories that encompass the diversity of complex ego-relative interaction patterns of the heterogeneous road objects in unstructured traffic conditions. We define 19 different explanation categories for the important objects: congestion, obstruction, on-road living being, stopped vehicle, avoid congestion, avoid obstruction, avoid on-road living being, avoid stopped vehicle, merging, cut-in, overtake, confrontation, crossing, slow down, deviate, left turn, right turn, u-turn, red light.

A bird’s eye view illustration for each of these explanations in different traffic situations is shown in Figure 1. Each row in the figure corresponds to a group of explanations with similar interaction patterns between the important road objects and the ego vehicle.

- Row A: Important objects appearing in the front of ego vehicle are in close interaction such that the ego vehicle’s driving path is blocked.
- Row B: The objects are relatively far such that sufficient driving space is available for the ego vehicle to avoid or move ahead of the object by deviating from its path.
- Row C: Lateral interactions between the ego vehicle and the important road objects.

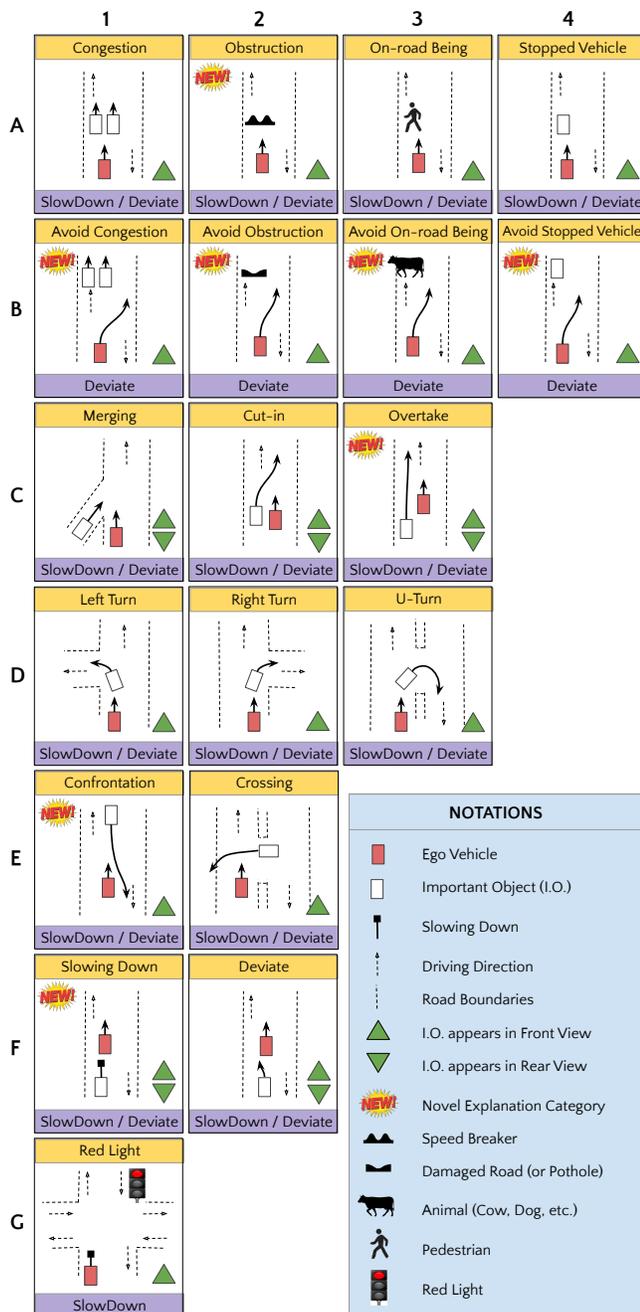


Fig. 1: Bird Eye View illustration of Explanations for Important Objects in different traffic situations captured in the IDD-X dataset. Each row (A to G) and column (1 to 4) pair is an explanation card. Inside each card, the yellow header corresponds to the explanation category while the purple footer corresponds to the possible ego vehicle’s driving behaviors in the demonstrated traffic situation. Icon notations used in the figure are defined in the bottom right corner. Important Object(s) in each card can either be visible in the front-camera view, in the rear-camera view, or in both of them. The corresponding notation for the views is shown in the bottom right corner of each card. The Driving Direction notation shows that the vehicles should ideally drive on their left-hand side of the road according to Indian traffic rules. The Novel Explanation Category icon means that the category is unique to this dataset, and has not been found in the existing driving datasets.

- Row D: Important objects doing a turning maneuver in front of the ego vehicle.
- Row E: Important objects approaching the ego vehicle such that the ego’s future trajectory is hindered.
- Row F: Primitive important object explanations whose consideration is crucial for the ego vehicle to make safe and careful driving decisions.

Broadly, Rows A, B, and G describe the passive influence of different road objects on ego vehicle’s driving decision. Rows C, D, E, and F define the ego-relative maneuvering styles of the important road objects. Note that the situations in Figure 1 are representative and variants may exist in data. For instance, in Row B, the ego vehicle may avoid the important objects from its left depending on the available driving space.

Samples of annotated driving scenarios from our dataset are shown in Figure 2. In scenario  $S_1$ , the ego vehicle slows down because multiple important road objects (motorcycle, truck, and autorickshaw appearing in front view  $S_1^F$ ) block ego vehicle’s driving path. This shows that ego vehicle’s driving behavior is influenced by the interactions of multiple road objects. Such scenarios are common in dense and unpredictable traffic conditions captured in our dataset.

The road objects appearing in the rearview may also impact the ego vehicle’s driving decision. In scenario  $S_1$ , the ego vehicle may collide with important objects (autorickshaw and motorcycle in rear view  $S_1^R$ ) if ego chooses to deviate from its driving path. The ego vehicle may collide with important object (car in rear view) if it chooses to abruptly slow down. Therefore, both the front and rear views are crucial for comprehensive analysis of the ego vehicle’s driving decisions.

#### A. Dataset Creation

We recorded 85 hours of dual-view driving videos from urban, rural, and highway roads in and around Hyderabad city, India. The dataset was captured in day, night, rainy, cloudy, and sunny climates and weather conditions. The front-view and the rear-view videos were captured with camera resolutions of  $2560 \times 1440$ , and  $1920 \times 1080$ , recording at 25 fps.

We used the open-source CVAT<sup>2</sup> tool for annotating the IDD-X dataset. The annotations are designed to capture all the important interactions of the surrounding road occupants when the ego vehicle takes a driving decision.

*Annotation procedure:* The annotators watch the driving videos and filter out video intervals or scenarios where the ego driver either slows down or deviates from the influence of any road entity. These driving behaviors are called stimulus-driven actions which were initially defined in the HDD dataset [6]. A total of 3635 driving scenarios were filtered out with their durations ranging from 0.3 to 11 seconds. These video intervals with annotated driving behaviors were analyzed by a different group of annotators (experienced drivers) to identify all the important road objects while imagining themselves driving the ego vehicle. The annotators create bounding boxes around such objects and track them by adjusting the bounding boxes across all frames at the frequency of 25Hz for the entire duration of the driving

<sup>2</sup><https://github.com/openvinotoolkit/cvat>

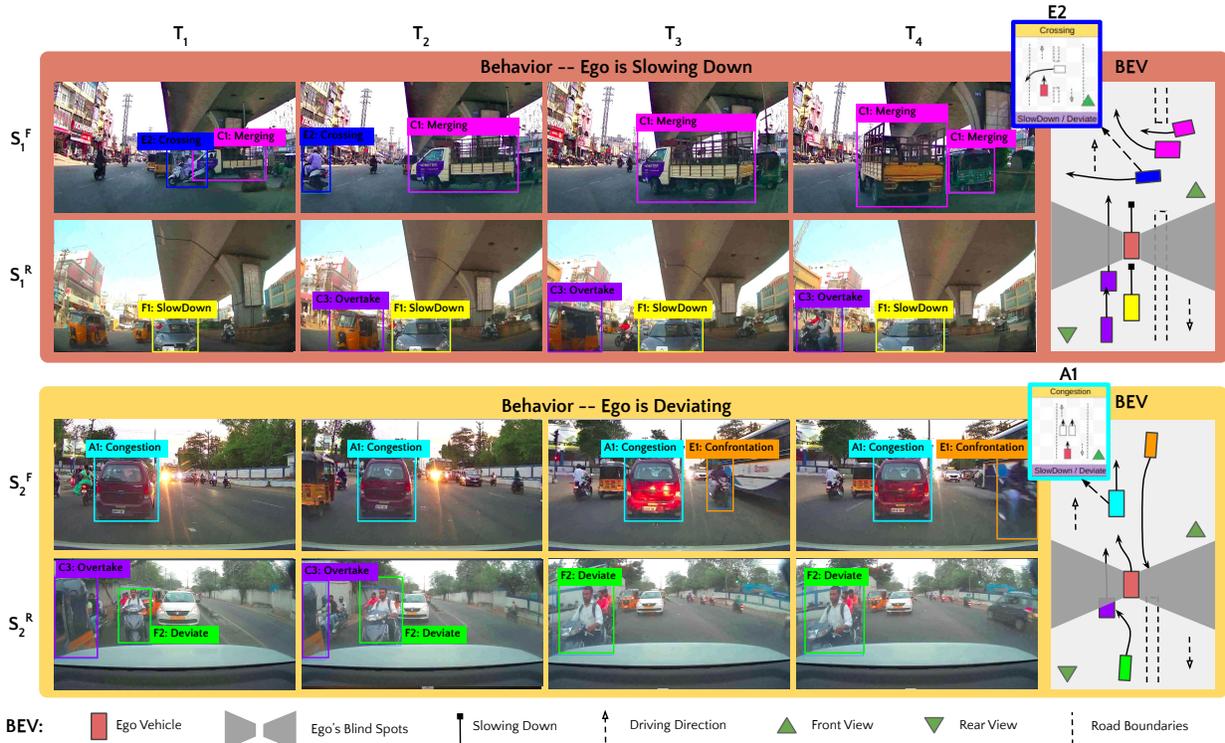


Fig. 2: Samples of the annotated driving scenarios in IDD-X. The ego vehicle’s driving behavior annotation is displayed at the top of each scenario ( $S_1$  and  $S_2$ ). The important object location annotations in front and rear views ( $S_1^F$ ,  $S_1^R$ ,  $S_2^F$ ,  $S_2^R$ ) at timeframes ( $T_1$  to  $T_4$ ) are shown with colored bounding boxes. Unique colors are assigned to the objects on the basis of their explanation category. The object’s explanation annotation is attached to each bounding box with the explanation card ID referenced from Figure 1. The Bird’s Eye View (BEV) illustration shows the trajectory of ego vehicle and annotated important objects as observed in four timeframes. The icon notations used in BEV are defined at the bottom of this figure. Sample explanation cards are shown in BEV for comparison of the important object’s maneuvering styles and traffic situations.

scenario. Next, they re-watch the entire video sequence, to specifically observe the interaction patterns of the important objects and assign suitable explanation labels from the pre-defined set of explanations described before. The annotators refer to the bird’s eye view illustration similar to Figure 1 for identifying the best-fitting explanation category for the road object. The explanation label and the object’s category information are provided as the object’s track attributes.

### B. Data Statistics

The dataset includes 3635 driving action scenarios containing 8689 important object annotations out of which 6427 appear in the front view and 2262 objects appear in the rear view. Figure 3 shows the distribution of the number of important objects observed per driving scenario. The larger number of observed important objects indicates denser traffic scenarios where 54% of the total scenarios in the dataset constitute more than one important object per scene. The distribution of explanations for heterogeneous road objects is shown in Figure 4. A heavy-tailed distribution of the labels can be observed in the figure.

## IV. TASKS

Given the inter-related nature of Important Object Localization and Important Object Explanation Prediction, we introduce a custom-designed multi-task deep network which obtains the related predictions for a given driving video.

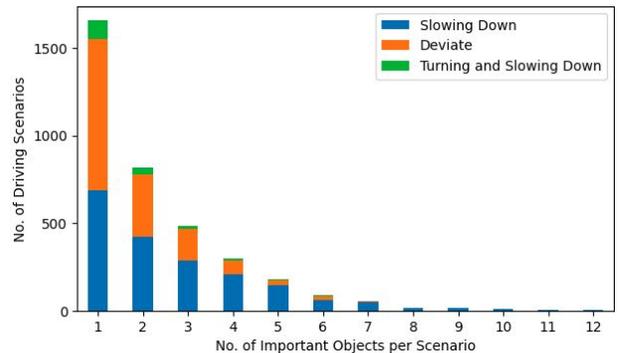


Fig. 3: Statistics of the number of Important Objects in a driving scenario for every driving action class.

Figure 5 illustrates the deep network along with essential components and procedures for the aforementioned tasks.

### A. Important Object Localization

In this task, we localize the important objects in a driving situation using their relative motion information. The approach for the task is divided into two steps 1) Multi-Object Tracking for localizing all the road objects observed in the driving scenario, and 2) Important Object Track Identification for predicting the importance of each object track.

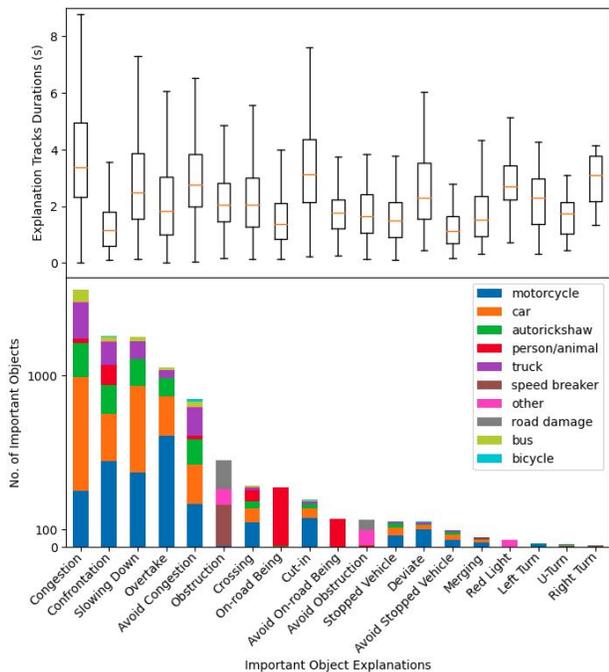


Fig. 4: Distribution of Explanations for Important Objects for every road object category and the average duration of the corresponding object tracks.

TABLE II: Driving Behavior Recognition. All scores are in %.

Modality	Slowdown	Deviate	Turn and Slowdown	Accuracy
RGB	79.4	57.3	12	65.5
Flow	<b>81.3</b>	<b>68.7</b>	<b>36</b>	<b>72.8</b>

**Multi-Object Tracking:** For this sub-task, we first detect all the objects in a driving scene observed from the front view camera and then use a tracking algorithm to track all the detected objects in subsequent video frames. A YOLOv4 detection model [17] was applied over all the frames in the front-view driving scenarios to get bounding boxes of all the road occupants. These bounding boxes were provided as input to the SORT tracker [18] algorithm for their association across frames to get their unique track-ids. This gives us the tracks of all the objects in the driving scene with their class labels.

**Important Object Track Identification:** This sub-task is formulated as a binary classification problem to predict the importance of an object track obtained from the tracking algorithm. We first prepare the training data with the binary importance labels for all the road objects in the driving scenes. Our dataset contains the ground-truth track information for the important road objects. We use this to label all the predicted object tracks, obtained from the previous sub-task, with the importance class. This is done by matching the predicted tracks with the ground truth tracks using spatial Intersection Over Union (IOU) between their bounding boxes averaged across all frames. The predicted object track is labeled as important if the IOU value is above a threshold otherwise it is marked as unimportant. This labeled training data is now used for importance prediction as described below.

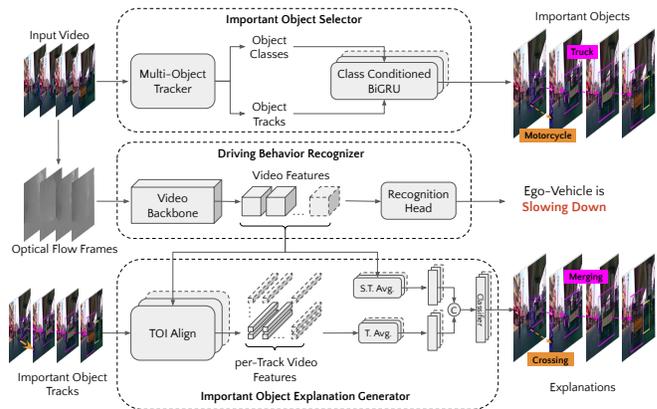


Fig. 5: The proposed approaches for Important Object Localization and Explanation in Dense and Unstructured Traffic.

The predicted object track’s bounding boxes and its object category are provided as input to an object class conditioned Bi-directional Gated Recurrent Unit (BiGRU) model [19] to output the importance class. A BiGRU model helps in effective temporal modeling of the object’s relative trajectory with respect to the ego vehicle. Additionally, it also considers the object category for differentiating between the bounding box size variations of different road objects. This is achieved by providing the encoded information of the object class along with its bounding box track coordinates at every time step inside the BiGRU model. Overall, our custom-designed approach helps in maintaining the relevant spatiotemporal context information throughout the track length. As shown in Figure 5, this model is separately applied over all the object tracks for their importance prediction.

TABLE III: Important Object Track Identification. All scores in %.

Model	Conditioning	Precision	Recall	F1-Score
BiGRU	-	35.0	86.1	49.8
Class Conditioned BiGRU	At $t = 0$	34.8	86.5	49.6
Class Conditioned BiGRU	At $t = T$	34.9	85.6	49.6
<b>Class Conditioned BiGRU</b>	<b>At all <math>t</math></b>	<b>35.8</b>	<b>86.8</b>	<b>50.7</b>

## B. Important Object Explanation

**Driving Behavior Recognition:** We adopt the TSN model [20] with Resnet-101 backbone for recognizing the driving behaviors in our dataset as it has demonstrated substantial success in video recognition tasks. The model was first pre-trained on the large-scale AVA action recognition dataset [21]. The pre-trained model was then used for transfer learning on our dataset with the trimmed video clips of the driving action scenarios. This is illustrated as the Driving Behavior Recognizer module in Figure 5, where the video backbone corresponds to the TSN model’s Resnet backbone with optical flow frames as input and the recognition head corresponds to the TSN’s average consensus head module for action prediction. The optical flow frames are extracted using the TVL1 algorithm [22] implemented in OpenCV.

**Important Object Track Explanation:** We use the ground truth important object tracks along with their explanation labels for training this task. The track bounding

TABLE IV: Important Object Explanation Prediction. F1-score for each explanation category is reported. All scores are in %.

Input Features	Congestion	Confrontation	Avoid Congestion	Overtaking	Crossing	Interfering Being	Cut-in	Avoid On-road Being	Avg. F1-Score	Wt. Avg. F1-Score
TOI-Aligned	72.2	48.6	53.3	53.9	37.5	17.4	7.3	0.0	36.3	54.8
TOI-Aligned + Context	71.0	53.6	47.2	48.6	44.7	22.4	34.7	25.0	43.4	55.8

boxes were resampled at the frame numbers corresponding to the extracted video features from the flow-based TSN model. The resampled bounding boxes were then used to extract the per-object track video features by applying the Track-Of-Interest Align (TOI-Align) operator [23] to the video features. Apart from this, the global driving action context information was also extracted by the spatial and temporal averaging of the video features from the flow-based TSN model. Finally, the global context information and the averaged per-track video features were concatenated and fed as input to a multi-layer perception (MLP) for explanation prediction. This is demonstrated as the Important Object Explanation Generator module in Figure 5, where it can be seen that multiple tracks video features are simultaneously extracted and fed to the classifier for explanation prediction.

## V. EXPERIMENTS AND RESULTS

We split the IDD-X dataset on the basis of driving action scenarios into train 70%, validation 15%, and test 15% sets. The number of ground truth important object annotations present in these driving scenarios comprised 6697 in training, 1728 in validation, and 264 in test. We consider the videos from the front view for the training and evaluation of all the tasks.

**Important Object Track Identification:** The dataset for this task was prepared using the YOLOv4 detection model [17] and the SORT tracking algorithm [18]. The YOLOv4 model was pre-trained on the COCO dataset [24] and fine-tuned on the IDD dataset [25] for detecting road objects specific to Indian driving scenarios. The tracks for each driving scenario were obtained from the SORT tracker algorithm. A standard threshold of 0.5 was used for the average spatial IOU value between the predicted and the ground-truth object tracks. A total of 10882 predicted important object tracks and 108042 unimportant object tracks were obtained.

This data was used for training the BiGRU model with a hidden size of 5, and a single hidden layer. Given the highly skewed nature between the important and the unimportant object classes, we adopted mini-batch class-balanced sampling strategy while training this model with standard cross-entropy loss. A batch size of 32 and a learning rate of 0.001 with Adam optimizer [26] were used for training the BiGRU model until 100 epochs. Table III shows the comparative performances of our model with different inputs. Utilizing the object class information at all timesteps enabled the best performance of 86.8% Recall and 50.7% F1-Score. Our experiments show that consideration of heterogeneous road object categories throughout the temporal domain enriches the model’s identification capability.

**Driving Behavior Recognition:** The input to the TSN model was randomly cropped video frames resized to 224x224, with random temporal sampling at the frame

interval of 1, clip length 5, and number of clips 8. The TSN model was trained with the standard cross-entropy loss, learning rate 0.000125, batch size 6, and for 40 epochs. The per-class recall and the weighted average accuracy of the predicted driving behaviors were used as evaluation metrics for this task. Table II shows the performances of the TSN driving action recognition models using different modalities: RGB and optical flow. The model trained with optical flow modality shows significantly better performance with a 7.3% improvement in overall accuracy. For specific driving actions, the per-class recalls also show superior performances with 1.9%, 11.4%, and 24% improvements in ‘Slowdown’, ‘Deviate’, and ‘Turn and Slowdown’ classes respectively. This finding is crucial and substantiates our choice to employ the optical flow-based TSN model for feature extraction in the explanation prediction task.

**Important Object Track Explanation:** The per-track video features were extracted using the trained TSN model with 256x224-sized image frames as input instead of 224x224. The frame width is modified so as to maintain the same aspect ratio with respect to the original image after resizing. This ensured unaltered aspect ratios for the track’s bounding boxes as well. The video features extracted from the last layer of the Resnet backbone in the TSN model with spatial dimension 7x13, temporal size 8, and channel size 2048. The track bounding boxes were downsampled to the video feature spatial dimensions before the application of the TOI-Align operator. The output from TOI-Align was 1x1x2048x8 (width x height x channel x time) per-object video features. The global context information obtained from spatial and temporal averaging of the video features was of dimension 1x1x2048x1. The concatenated output, using the global context information and the per-object track features, was of dimension 4096. This was fed as input to an MLP with a hidden layer of size 128. The explanation labels in the training dataset with a total count of less than 100 were not considered for the training and evaluation of this task. Standard cross-entropy loss with learning rate 0.001, and batch size 16, was used for training the MLP classifier until epochs 50. Per-class F1-score, average F1-score, and weighted average F1-score are reported as evaluation metrics for this task.

We compare two variants: TOI-Aligned per-track features, and TOI-Aligned per-track features combined with global context information. As observed from the Table IV, the TOI-Aligned model excels in the ‘Congestion’, ‘Avoid Congestion’, and ‘Overtaking’ categories, with better per-class F1 scores. When the global context information is added, the performance of the tail classes (‘Crossing’, ‘Interfering Being’, ‘Cut-in’, and ‘Avoid On-road Being’) improves by a large margin (7.1% for average F1-score, 1% for weighted average F1-score). Overall, the results suggest that inclu-

sion of global context information significantly improves the model's ability to correctly identify the underprivileged explanation categories. However, for more common driving situations (e.g. 'Congestion' and 'Overtaking') per-track-features-only appear to suffice. The performance in categories such as 'Cut-in' and 'Avoid On-road Animal' is low across both variants, pointing out areas for improvement.

**Acknowledgement** The project is funded by the iHubData and Mobility at IIIT Hyderabad. We thank the data collection and annotation team for their effort.

## REFERENCES

- [1] J. Li, H. Gang, H. Ma, M. Tomizuka, and C. Choi, "Important object identification with semi-supervised learning for autonomous driving," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2913–2919.
- [2] M. Gao, A. Tawari, and S. Martin, "Goal-oriented object importance estimation in on-road driving videos," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5509–5515.
- [3] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1043–1052.
- [4] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos, "Explainable object-induced action decision for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9523–9532.
- [5] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 563–578.
- [6] V. Ramanishka, Y.-T. Chen, T. Misu, and K. 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*, 2018, pp. 7699–7707.
- [7] R. Chandra, X. Wang, M. Mahajan, R. Kala, R. Palugulla, C. Naidu, A. Jain, and D. Manocha, "Meteor: A dense, heterogeneous, and unstructured traffic dataset with rare behaviors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9169–9175.
- [8] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.
- [9] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2942–2950.
- [10] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale navigation video datasets," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2174–2182.
- [11] T. Wu, A. Luo, R. Huang, H. Cheng, and Y. Zhao, "End-to-end driving model for steering control of autonomous vehicles with future spatiotemporal features," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 950–955.
- [12] H. Ben-Younes, É. Zablocki, P. Pérez, and M. Cord, "Driving behavior explanation with multi-level fusion," *Pattern Recognition*, vol. 123, p. 108421, 2022.
- [13] T. Jing, H. Xia, R. Tian, H. Ding, X. Luo, J. Domeyer, R. Sherony, and Z. Ding, "Inaction: Interpretable action decision making for autonomous driving," in *European Conference on Computer Vision*. Springer, 2022, pp. 370–387.
- [14] C. Li, Y. Meng, S. H. Chan, and Y.-T. Chen, "Learning 3d-aware ego-centric spatial-temporal interaction via graph convolutional networks," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8418–8424.
- [15] Y. Yao, X. Wang, M. Xu, Z. Pu, E. Atkins, and D. Crandall, "When, where, and what? a new dataset for anomaly detection in driving videos," *arXiv preprint arXiv:2004.03044*, 2020.
- [16] S. Dokania, A. Hafez, A. Subramanian, M. Chandraker, and C. Jawahar, "Idd-3d: Indian driving dataset for 3d unstructured road scenes," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4482–4491.
- [17] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [18] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.
- [19] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [20] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.
- [21] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, *et al.*, "Ava: A video dataset of spatio-temporally localized atomic visual actions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6047–6056.
- [22] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l1 optical flow," in *Pattern Recognition: 29th DAGM Symposium, Heidelberg, Germany, September 12-14, 2007. Proceedings 29*. Springer, 2007, pp. 214–223.
- [23] G. Singh, V. Choutas, S. Saha, F. Yu, and L. Van Gool, "Spatio-temporal action detection under large motion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6009–6018.
- [24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
- [25] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. Jawahar, "IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1743–1751.
- [26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.