DRAMA: Joint Risk Localization and Captioning in Driving - Supplementary Material

Srikanth Malla Chiho Choi Isht Dwivedi Joon Hee Choi Jiachen Li

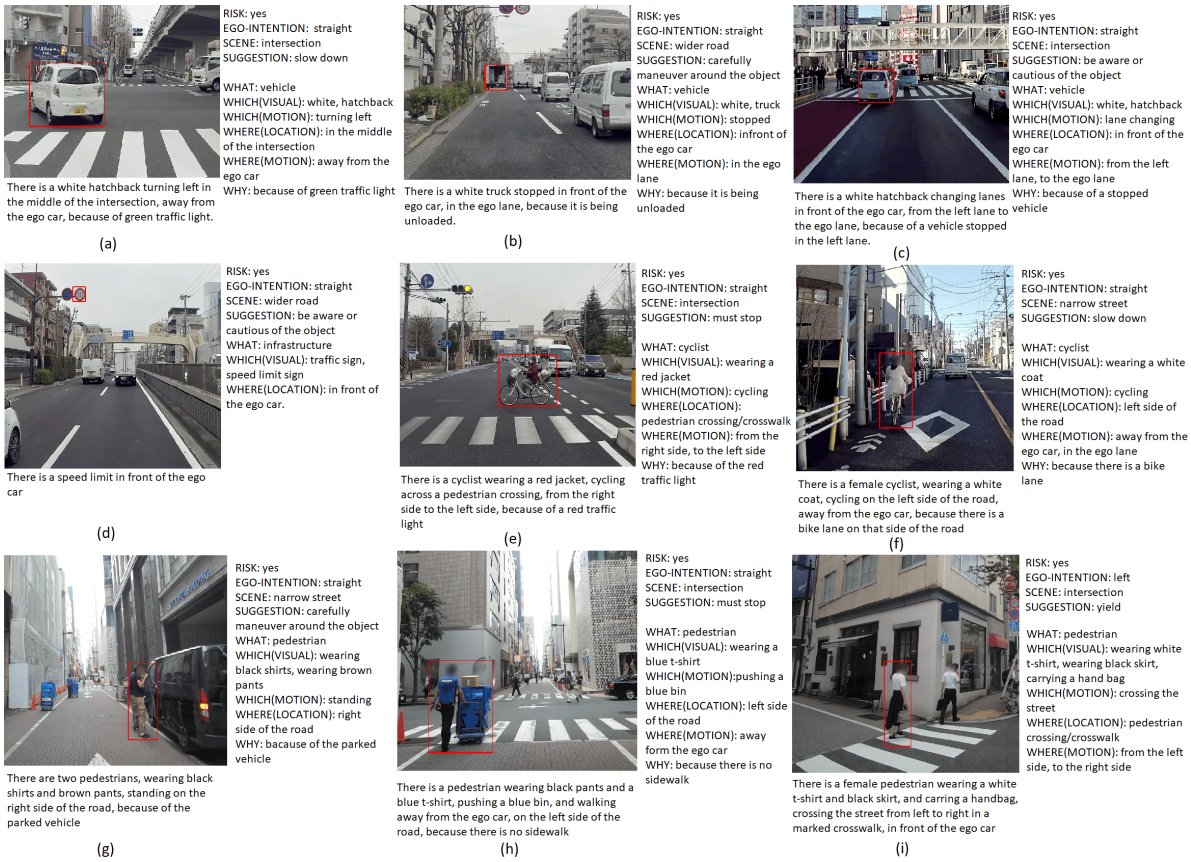


Figure 8: Example scenarios of the DRAMA dataset. Risks are perceived while interacting with different vehicles in (a-c), infrastructure in (d), cyclists (e,f), and pedestrians(g-i).

A. Additional Analysis of DRAMA

A.1. Example Scenarios

We visual various scenarios of our DRAMA dataset in Figure 8. Each scenario includes video- and agent-level attributes annotated from sequence and motion observations. Some actions of vehicles such as turning left, stopped, lane changing are shown in Figure 8a-8c respectively in various interactive environments (intersection, wide and narrow road) where the reasoning of the important object is described by 'because of traffic light' in 8a, 'because it is being unloaded' in 8b, ' because of a stopped vehicle' in 8c, respectively. We show cyclists interacting with the ego car at 'intersections or crosswalks' in 8e and 'narrow roads or driving in ego lane' 8f. Pedestrians are also identified as important agents interacting with the ego-car while performing different actions such as 'standing', 'pushing','crossing the street' as shown in Figure 8g-8i. Their importance is reasoned by behaviors like 'because of parked vehicle'

	Object-level Q/A Attributes					Bounding Box		Captioning	
Class (what)	visual attr (which)	motion attr (which)	location (where)	motion direc (where)	ext fact (why)	no of boxes	normalized mean box size	unique words	total occurrences
Veh	24,546	13,829	12,350	10,462	14,712	12,273	0.0846	533	214,511
Ped/Cyc	7,338	3,829	3,454	3,571	4,080	3,344	0.0405	608	75,576
Inf	3,154	-	1,505	-	-	1,449	0.0102	268	16,616
All	35,038	17,658	17,309	14,033	18,792	17,066	0.0697	992	306,708

Table 4: The statistics of obect-level Q/A attributes, bounding box, and captions.

in 8g and 'because there is no side walk' in 8h. In case of Figure 8i, the risk is caused by internal stimuli rather than affected by external influences, so WHY is not annotated in this scenario. Similarly, the infrastructure only has visual and location questions in Figure 8d.

A.2. Statistics

We show the data statistics of DRAMA in Table 4. Note that video-level Q/A attributes are not included in the table as the annotated number is same as the number of scenarios (17,785).

The visual attributes for vehicles are 2 per a bounding box on average (PBA) as they can be described by the color and type of the vehicle. Whereas, pedestrians/cyclists has 2.19 PBA, which indicates these agents have more descriptive representations. The visual attributes for the infrastructure usually include it's state and name, PBA is measured at 2.17. The motion attributes for vehicles is 1.12 PBA, as it describes the motion state like stopped, parked, etc. Similarly for pedestrians/cyclists, it is 1.14 PBA. The location attributes for vehicles, pedestrians/cyclists, and infrastructure are respectively 1.00, 1.03, 1.06 PBA. The motion direction for pedestrian/cyclists is 1.06 PBA, which is higher than that of vehicles (0.85 PBA). This is mainly because most of the influence happens when pedestrians/cyclists cross the road or cyclists move slowly away/towards the ego-car, which requires more descriptions. On the other hand, in case of vehicles, they are identified important when they slow down (or braking) in front of the ego-vehicle without describing the direction. The external factors or second level of reasoning (reasoning for the important agent's behavior) for vehicles is 1.20 PBA and pedestrians/cyclists 1.22 PBA. The infrastructure neither moves nor gets influenced by other external factors, so the last two columns in agent level VQA attributes in Table 4.

The bounding box proportions across different object categories are also shown in Table 4. The vehicle bounding boxes are the most dominant ones 12,273 (71.91%) of the all the boxes. The normalized mean bounding boxes size (NMBS) of the vehicle category is 8% from the entire image as they are often very close to the ego-vehicle. There are 2,909 pedestrians and 435 cyclists, which is 19.19% of all the bounding boxes in the DRAMA dataset. The NMBS of this object category is 4% of the image. We can infer that the ego-car slows or stops from a distant location when their sizes are small from the egocentric perspective. The bounding boxes of infrastructure are the least percentage 1,449 (8.49%), with NMBS of 1.02%. Most of infrastructure is construction cones, traffic lights, and traffic signs, which cover only a small portion of the image. Resulting in poorer performance for object localization for infrastructure compared to others in Table 3 can be partially described from their NMBS.

In total, the vocabulary used in the dataset consists of 992 unique words, with total occurrences of 306,708. The vocabulary size of vehicles is 533, with total occurrences of 214,511 (69.94%). The vocabulary size of pedestrians and cyclists is 608 with total occurrences of 75,576 (24.64%). The vocabulary size of infrastructure is 268 with total occurrences of 16,616 (5.41%). The proportions of the words are proportional to number of those objects present in the dataset. The common unique words used in captioning across all classes are 112 with total occurrences of 265,922 (86.7%). Additional information can be found in the main manuscript in Section 3.2 and Figure 4.

A.3. Privacy

An open-source tool, Anonymizer [37] is used to anonymize faces and vehicle number plates as a coarse annotation step. The annotators find unblurred faces or license plates and manually blurred them.

A.4. Question Representation

For a comprehensive annotation of DRAMA, we generate a set of questions that can yield various forms of answers from closed-ended such as boolean (yes or no) to open-ended such as 'pedestrian wearing black pants and white shirt' or 'orange traffic cones on the right'. We used elementary operations (*i.e.*, what, which, where, why, and how) to reason about ego-driver's behavioral response to perceived risk that activates braking of the vehicle. The questions for querying

	Layer	Kernal shape	Output shape
Flow/Visual Encoder			
0	resnet.Conv2d_0	[3, 64, 7, 7]	[1, 64, 370, 500]
1	resnet.BatchNorm2d_1	[64]	[1, 64, 370, 500]
2	resnet.ReLU_2	-	[1, 64, 370, 500]
3	resnet.MaxPool2d_3	-	[1, 64, 185, 250]
..	... (resnet layers)		
306	resnet.7.2.Conv2d_conv3	[512, 2048, 1, 1]	[1, 2048, 24, 32]
307	resnet.7.2.BatchNorm2d_bn3	[2048]	[1, 2048, 24, 32]
308	resnet.7.2.ReLU_relu	-	[1, 2048, 24, 32]
309	AdaptiveAvgPool2d_adaptive_pool	-	[1, 2048, 14, 14]
Decoder			
310(a)	concat.visual_flow	-	[1, 4096, 14, 14]
310(b)	module.Embedding_embedding	[512, W]	[1, T, 512]
311(a)	Linear_init_h	[4096, 512]	[1, 512]
311(b)	Linear_init_c	[4096, 512]	[1, 512]
312	LSTM_cell_with_attention	-	[1, 512]
313	word_dropout_0	-	[1, 512]
314	word_fc	[512,W]	[1, W]
315	word_softmax	-	[1, W]
316	rollout LSTM for M times, from step 312 to 315		
317	Self-attention Step, Sec 4		[1, 512]
318	bbox_predictor.Linear_0	[512, 4]	[1, 4]
319	bbox_predictor.Sigmoid_1	-	[1,4]

Table 5: DRAMA summary, for input image of size (3,740,1000), batch size as 1, vocabulary size (W) as 989 and max sentence length (M) as 50

video- and object-level attributes are as follows.

Video-level questions:

- Is there any '*risk*' in the scene?
- What is the '*intention*' of the ego-car?
- What is this '*scene*'?
- What '*suggestions*' do you give for the driver to avoid risk?

If risks are perceived in the video, questions regarding the important agent are:

- '*What*' agent is it?
- If Pedestrian/Cyclist, '*Which*' visual attributes best describe the important agent?
- If Vehicles, '*Which*' type of the vehicle is it? (visual attributes)
- If Vehicles, '*Which*' color is the vehicle? (visual attributes)
- If Infrastructure, '*Which*' name is the infrastructure called? (visual attributes)
- If Infrastructure, '*Which*' state is it in? (visual attributes)
- '*Which*' motion attributes best describe the agent?
- '*Where*' is the location of the agent?
- '*Where*' is the agent moving (from or to)?
- '*Why*' is the object behaving the way it is?

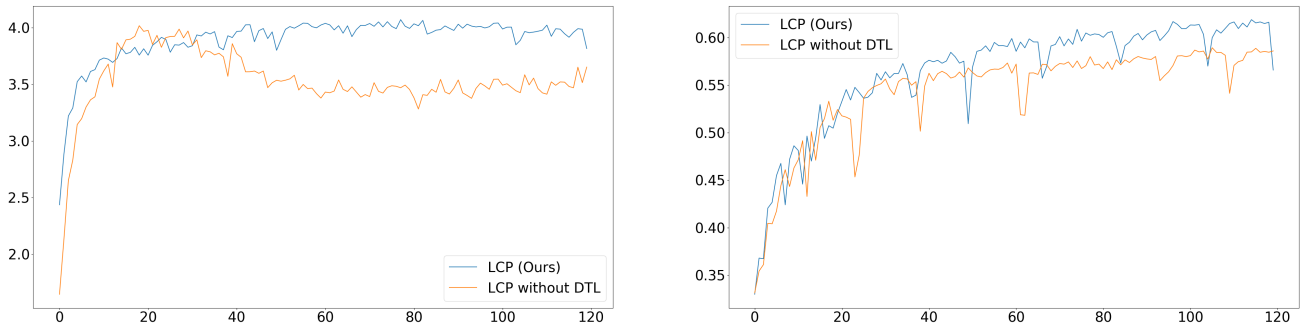
B. Implementation Details

B.1. Bounding Box Parameterization

The bounding box is notated using {cx,cy,l,w}, where (cx,cy) is the center coordinate and (l,w) are the dimension of the bounding box. They are normalized using the size of the image (L,W). During inference, we use a sigmoid activation function to find the values in the range of [0,1] and project it back to the original pixel coordinates to evaluate the output using the metrics (*i.e.*, Mean-IOU and Accuracy).

B.2. Network Architecture

The network architecture is shown in Table 5. Both visual encoder and flow encoder use ResNet-101 to extract features using the layers 0-309. Layer 0-308 is taken from original ResNet-101, and we add Layer 309 to generate features with a fixed size of (2048,14,14). In the decoder, Layer 310(a) concatenate both visual



(a) (b)

Figure 9: (a) CIDEr score on validation set during training and (b) Mean-IOU score on validation set during training.

Metrics	B1	B4	M	R	C	S	Mean-IOU	Acc(IOU¿0.5)
LCP (Ours)	0.739±0.0065	0.547±0.011	0.391±0.0045	0.700±0.004	3.724±0.0765	0.560 ±0.009	0.614±0.0075	0.684±0.0115

Table 6: Table showing standard deviations of our model for each metrics

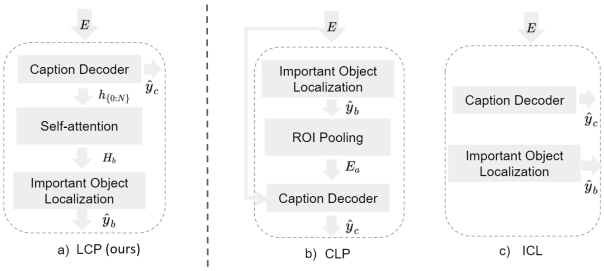


Figure 10: Comparison of our decoder (LCP) with that of two baselines (CLP and ICL).

and flow features and the output size is (4096,14,14). Layer 310(b) is to convert the word representations from LongTensor to embeddings with a size of 512 with T as maximum sentence length 50 (after padding zeros). Layer 311(a,b) converts the concatenated embedding to initial hidden state (a) and cell state (b) for the LSTM, followed by the cell update in Layer 312, using attention mechanism on the encoded image features, please refer to SAT[30] for more details. Layer 313 (dropout), 314 (fully connected), and 315 (softmax) is the conversion of LSTM hidden state to the output word at each time step. We rollout for T times through Layer 312-315. In practice, to save computation time each sentence is rolled out only till the length of the ground truth sentence during training. In Layer 317, we use the self-attention mechanism explained in the main draft Sec 4 (sub-section decoder), which takes all the hidden states of the LSTM and performs the self-attention operation. The output from self attention is converted to bounding box parameters from Layer 318-319 using fully connected layer (Layer 318) and Sigmoid operation (Layer 319).

B.3. Block Diagrams

In Figure 10, we visualize the block diagram of decoders used for evaluation. We refer to Section 4 and 5 in the main manuscript for overview ours (LCP) as well as other baseline models (CLP and ICL). Details of ROI pooling used in CLP is presented in Fast-RCNN [36].

B.4. Model Training

We build our framework using the PyTorch framework with Tesla V100-SXM2-32GB GPUs. The batch size is set to 16, and we use Adam optimizer with a learning rate of 1e-4 for the encoder and 4e-4 for the decoder. The dropout ratio is 0.5 in the word_dropout_0 layer in Table 5. The model is trained for 120 epochs, and we report the best checkpoint (*i.e.*, saved using the CIDEr score measured from the validation set) to evaluate the model on the test set.

C. Additional Evaluation

C.1. Model Robustness

We ran our best model for three times and computed the standard deviation for the metrics on the test set. As shown in Table 6, the small value of standard deviation demonstrates that our approach consistently provides robust prediction capabilities.

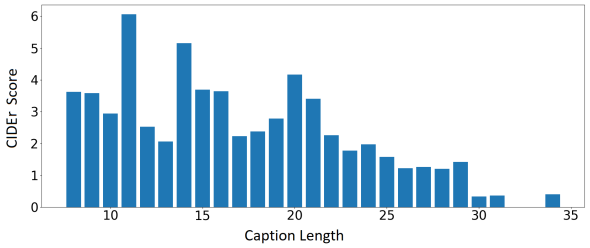


Figure 11: Caption Length vs Average CIDEr Score are evaluated on the test set.

C.2. Caption Length vs CIDEr Score

In Figure 11, we plot the CIDEr score of our model with respect to the length of generated captions using the test set. The CIDEr score is sorted in the order of ascending length. In general the performance degrades with the increase in the length or complexity of the sentence. The increase in performance at several lengths is because the captions are more commonly observed with those lengths.

C.3. Multi-Task Training

We compare two metrics, CIDEr and Mean-IOU, while training our model with (LCP Ours) and without (LCP without DTL) multi-task weighting. The captioning metric CIDEr performance is shown in Figure 9a. The important object identification metric Mean-IOU performance is shown in Figure 9b. We weigh the model 'LCP without DTL' with $\lambda = 15$ in Equation 4, which we found the best performing model from the manual weighting. The model 'LCP (Ours)' is trained using the loss in Equation 3. L_c, L_b refers to the caption loss and bounding box loss as mentioned in Section 4 of the main manuscript. As shown in Figure 9, 'LCP (Ours)' shows the continuous improvement of the performance for both tasks. However, 'LCP without DTL' sacrifices the capability of one task (captioning) while improving the other (object identification).

$$L_{sum} = L_c + \lambda * L_b \tag{4}$$

C.4. Qualitative Analysis

In Figure 12, we visualize additional successful cases in addition to Figure 7 shown in the main manuscript. 12a and 12c are two different pedestrians walking on a narrow road and at the intersection. In both cases, the visual, location, motion direction attributes are generated correctly. Similarly, the cyclist in 12b is identified as an important agent with a good reasoning ('because there is no bicycle lane'). In 12d, the traffic light is predicted correctly even when there is a vehicle in front of the ego car.

In Figure 13, we show the captioning failures while the localization of the important agent is correct. In 13a, the truck is found correctly but its reasoning is incomplete. In 13b, the network reasoned about stopping as red traffic light. The model might be confused because the traffic congestion is hard to see in this scenario. In contrast, 13c is an example of the opposite case, where our model is hard to see the traffic light as a reason for stopping. In 13d, the cyclist's motion direction is predicted wrong which resulted in incomplete reasoning.

In Figure 14, localization failures are shown. The network sometimes gets confused to identify the important object when there are multiple objects shown reasonably important in the scene. We show several cases: '*vehicle→pedestrian*' *in 14a*, '*pedestrian→vehicle*' *in 14b*, '*infrastructure→pedestrian*' *in 14c*, '*infrastructure→ infrastructure*' *in 14d* . In scenarios of 14a-c, the identified important object is rather closer to the ego-vehicle whereas the ground truth is annotated very far. The encoder of our model might be hard to extract image features of those agents (as their size is very small). The advancement of the algorithm to cover such cases are our future research direction. In scenario 14d, there exist multiple traffic lights showing the same traffic information. Although our identification is different from the ground truth, the model successfully reasons about the interaction as a caption, which activated braking of the vehicle.

D. Limitations

There has been a significant progress in computer vision and machine learning algorithms for autonomous and cooperative driving. However, explainable models that ground language and vision in this domain is less highlighted. In this work, we introduce a new dataset and address risk localization with its reasoning as a language description in driving scenarios where such a model can benefit to situational awareness in autonomous driving and driving assistant systems. Our setting of interest and the dataset have an objective of addressing driving risks that implicitly and explicitly influence safety-critical design of intelligent systems. Therefore, real-world applications that adopt such reasoning models may cause injury or death to personnel or loss of property if they fail to make appropriate predictions. Although we provided several failure cases to better understand the limitation of the models and our problem settings, the users should be aware of other failures not explored in this paper. Besides the model, our dataset is labeled by humans, which may incorporate biases due to errors in interpretation. The users are also requested to assess the risk while using our

Figure 12: Successful cases. Ground truth in green color and Prediction in red color.

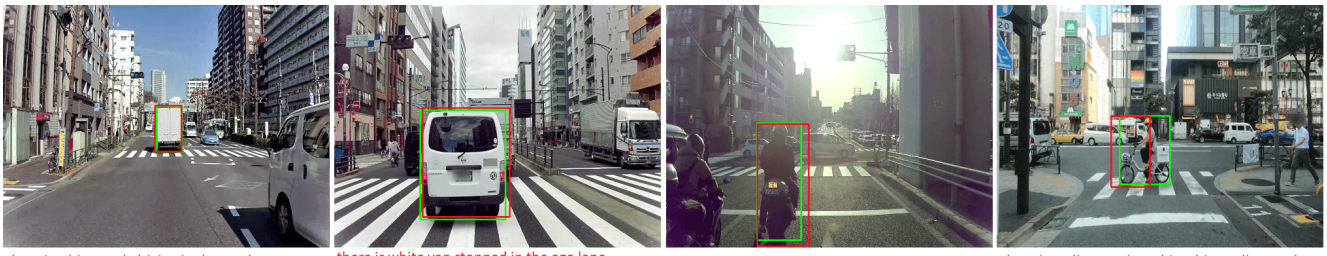


Figure 13: Captioning failures. Ground truth in green color and Prediction in red color.

Figure 14: Localization fail cases. Ground truth in green color and Prediction in red color.

dataset for their safety critical applications.

References

[36] Ross Girshick. Fast r-cnn. InProceedings of the IEEE international conference on computer vision, pages1541440–1448, 2015

[37] https://github.com/understand-ai/anonymizer