TITAN: Future Forecast using Action Priors
- Supplementary Material

Srikanth Malla Behzad Dariush Chiho Choi
Honda Research Institute USA
{smalla, bdariush, cchoi}@honda-ri.com

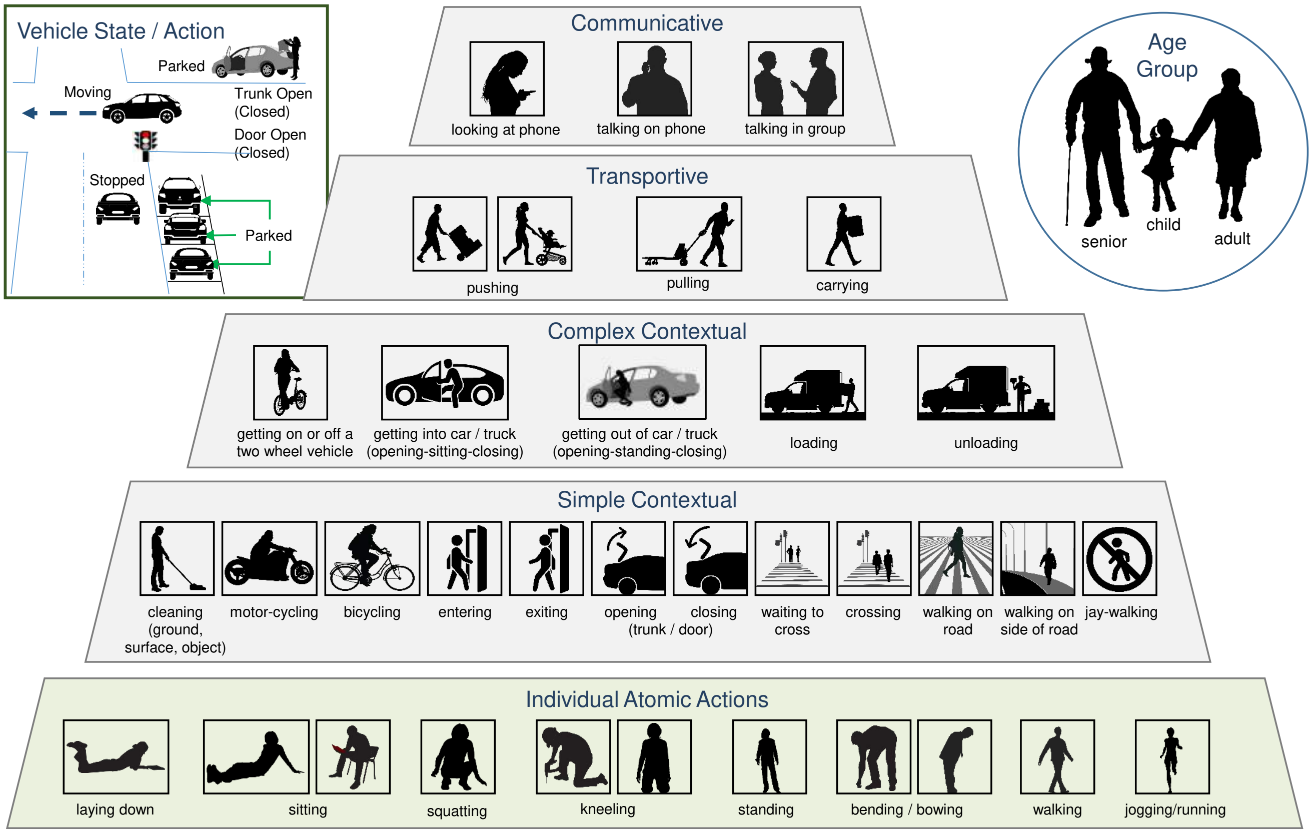


Figure 1: Our TITAN dataset contains 50 labels including vehicle states and actions, pedestrian age groups, and targeted pedestrian action attributes that are organized hierarchically corresponding to atomic, simple/complex-contextual, transportive, and communicative actions.

A. Details of the TITAN Dataset

Figure 1 illustrates the labels of the TITAN dataset, which are typically observed from on-board vehicles in driving scenes. We define 50 labels including vehicle states and actions, pedestrian age groups, and targeted pedestrian action attributes that are hierarchically organized from primitive atomic actions to complicated contextual activities. Table 1 further details the number of labels, instances, and descriptions for each action set in the TITAN dataset. For pedestrians, we categorize human actions into 5 subcategories based on their complexities and compositions. Moreover, we annotate vehicle states with 3 sub-categories of motion, and trunk / door status. Note that the trunk and

Category	Set	# Classes	# Instances	Description
Human Action	Atomic	9	392511	atomic whole body actions/postures that describe primitive action poses (*e.g.*, sitting, standing, walking, etc.)
	Simple contextual	13	328337	single atomic actions that include scene context (*e.g.*, jaywalking, waiting to cross)
	Complex contextual	7	5084	a sequence of atomic actions with increased complexity and/or higher contextual understanding
	Transportive	4	35160	manually transporting an object by carrying, pulling, or pushing.
	Communicative	4	57030	communicative actions (e.g. talking on the phone, looking at phone, or talking in groups.)
Vehicle State	Motion status	3	249080	motion status of 2-wheeled and 4-wheeled vehicles (parked / moving / stationary)
	Trunk status	2	146839	trunk for 4-wheeled vehicles (open / closed)
	Door status	2	146839	door status for 4-wheeled vehicles (open / closed)
Other Labels	Age group	3	395769	subjective categorization of pedestrians into age groups (child / adult / senior)
	Object type	3	645384	participant types categorized into pedestrian / 2-wheeled / 4- wheeled vehicles

Table 1: Details of the TITAN dataset. We report the number of labels, instances, and descriptions for each action set.

door status is only annotated for 4-wheeled vehicles. Vehicles with 3-wheels without trunk but with doors are annotated as 4-wheeled and trunk open. Also, 3-wheeled vehicles with no trunk and doors are annotated as 2-wheeled vehicles. The list of classes for human actions is shown in Table 2. The annotators were instructed to only localize pedestrians and vehicles with a minimum bounding box size of 70×10 $pixels$ and 50×10 $pixels$ in the image, respectively.

Several example scenarios of TITAN are depicted in Figure 2. In each scenario, four frames are displayed with a bounding box around a road agent. We also provide their actions below each frame. Note that only one agent per frame is selected for the purpose of visualization. The same color code is used for each action label, which can be found in Figure 2 of the main manuscript.

B. Additional Evaluation

In this section, we provide additional evaluation results of the proposed approach.

B.1. Per-Class Quantitative Results

In Table 2, we present per-class quantitative results of the proposed approach, which are evaluated using the test set of TITAN. Note that the number of instances for some actions (*e.g.*, *kneeling*, *jumping*, etc.) are zero, although they are present in the training and validation set. This is because we randomly split 700 clips of TITAN into training, validation, and test set. We will regularly update TITAN to add more clips with such actions.

We observe that the error rate for some classes are either much lower or higher than other classes. For example, scenarios depicting *getting into a 4 wheel vehicle*, *getting out of a 4 wheel vehicle*, and *getting on a 2 wheel vehicle* show very small FDE as compared to others. Also, scenarios depicting *entering a building* has a larger ADE and FDE than other scenarios. The reason for this can be explained by considering interactions of agents. When a person is *getting into a vehicle*, the proposed interaction encoder builds a pair-wise interaction between the person (subject that the action generates) and the vehicle (object that the subject is related to). It further validates the efficacy of our interaction encoding capability. In contrast, no interactive object is given to the agent for *entering a building* class since we assume agents are either pedestrians or vehicles. As mentioned in the main manuscript, we plan to incorporate additional scene context such as topology or semantic information.

B.2. Efficacy of Multi-Task Loss

The comparative results of the I3D action recognition module with and without the multi-task (MT) loss is shown in Table 3. The performance improvement for atomic and simple contextual actions for pedestrians and motion status for vehicles with the MT loss validates its efficacy of modeling aleatoric homoscedastic uncertainty of different tasks.

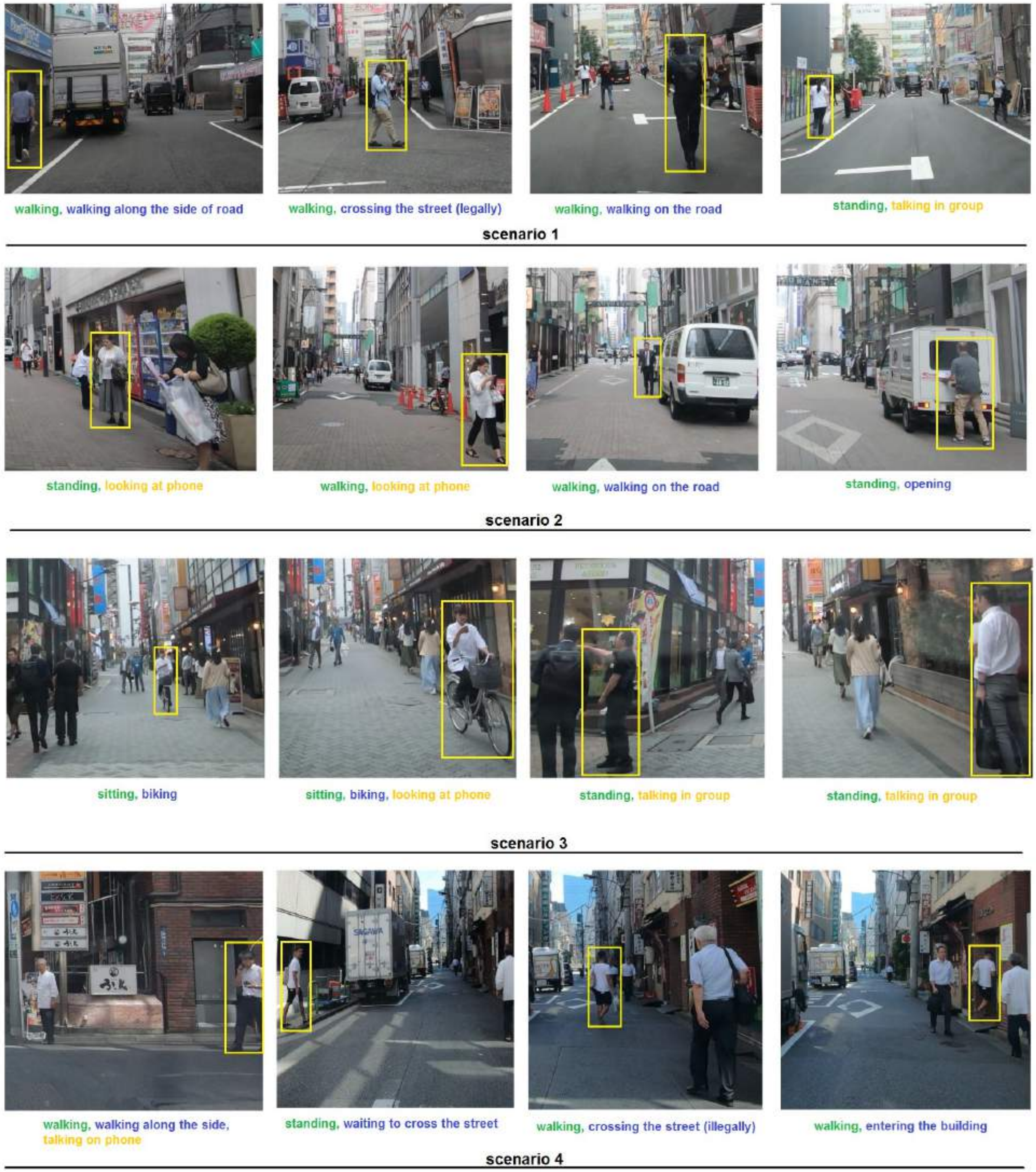
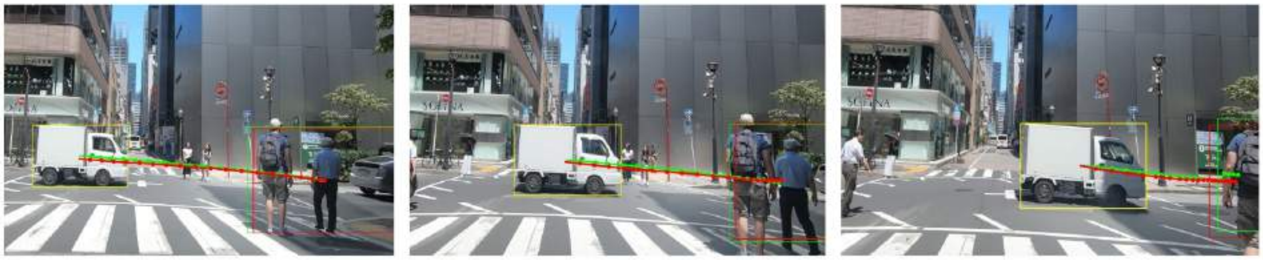

walking, walking along the side of road
walking, crossing the street (legally)
walking, walking on the road
standing, talking in group
scenario 1
standing, looking at phone
walking, looking at phone
walking, walking on the road
standing, opening
scenario 2
sitting, biking
sitting, biking, looking at phone
standing, talking in group
standing, talking in group
scenario 3
walking, walking along the side, talking on phone
standing, waiting to cross the street
walking, crossing the street (illegally)
walking, entering the building
scenario 4

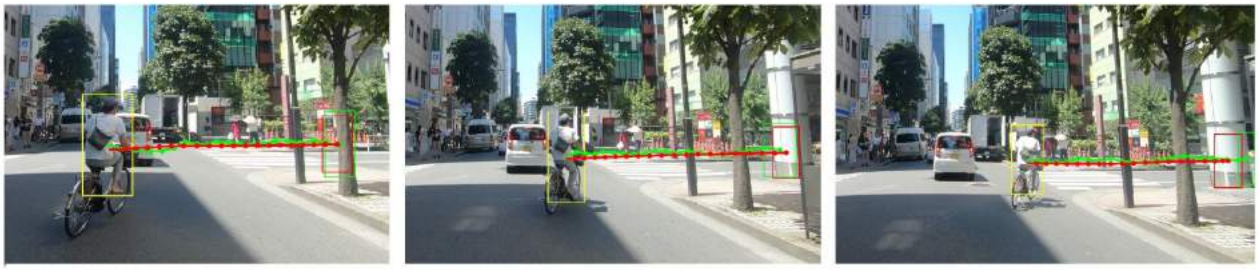
Figure 2: Example sequences from the TITAN dataset. Some notable actions are highlighted with different color codes following the hierarchy in the main manuscript (Color codes: Green - atomic, Blue - simple contextual, and Yellow - communicative). Images are cropped from their original size for better visibility.

Action Set	Class	ADE↓	FDE↓	FIOU↑	#Instances
Atomic Action	standing	10.56	18.63	0.6128	1476
	running	12.39	19.95	0.6179	89
	bending	12.76	20.85	0.6560	156
	kneeling	0.00	0.00	0.00	0
	walking	13.31	23.15	0.5712	6354
	sitting	11.10	20.74	0.6282	337
	squatting	11.90	18.82	0.5598	4
	jumping	0.00	0.00	0.00	0
	laying down	0.00	0.00	0.00	0
	none of the above	9.69	16.43	0.7408	7237
Simple-Contextual	crossing at pedestrian crossing	13.22	21.59	0.5976	881
	jaywalking	13.10	21.91	0.6148	340
	waiting to cross street	11.49	21.75	0.5783	65
	motorcycling	20.00	31.81	0.5494	4
	biking	13.22	21.13	0.6283	287
	walking along the side of the road	11.33	24.50	0.5516	2668
	walking on the road	13.41	22.30	0.5794	2486
	cleaning (ground, surface, object)	11.67	22.58	0.6502	19
	closing	9.84	20.50	0.4947	14
	opening	12.99	29.89	0.1995	13
	exiting a building	13.56	28.09	0.5264	61
	entering a building	28.06	53.02	0.2259	6
	none of the above	9.85	16.76	0.7201	8809
Complex-Contextual	unloading	11.07	18.45	0.7082	37
	loading	11.59	18.54	0.6652	40
	getting in 4 wheel vehicle	8.39	10.80	0.5682	10
	getting out of 4 wheel vehicle	9.63	9.58	0.7972	3
	getting on 2 wheel vehicle	7.73	11.16	0.7619	10
	getting off 2 wheel vehicle	0	0	0	0
	none of the above	11.32	19.54	0.6557	15553
Communicative	looking at phone	12.12	21.48	0.6435	392
	talking on phone	11.69	19.39	0.6056	268
	talking in group	11.70	20.82	0.6025	461
	none of the above	11.28	19.43	0.6588	14532
Transportive	pushing	12.57	23.07	0.6148	232
	carrying with both hands	11.39	20.23	0.6477	445
	pulling	12.01	21.29	0.5198	76
	none of the above	11.29	19.44	0.6574	14900
Motion-Status	stopped	8.96	23.08	0.6148	232
	moving	9.18	20.23	0.6477	445
	parked	9.93	21.29	0.5199	76
	none of the above	12.72	19.44	0.6574	14900

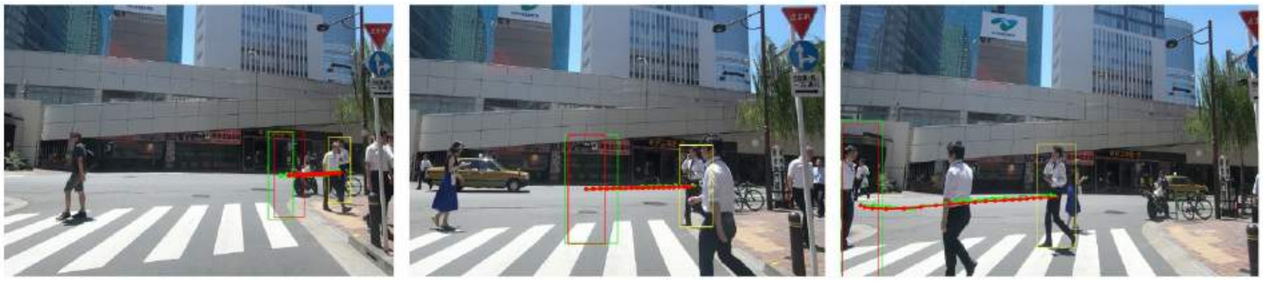
Table 2: Per-class evaluation results using the test set of 100 clips.



Scenario 5



Scenario 6



Scenario 7

Figure 3: Qualitative results of TITAN from different sequences. Trajectories in Red denote predictions, trajectories in green color denote ground truth, and a yellow bounding box denotes the last observations. (Images cropped for better visibility)

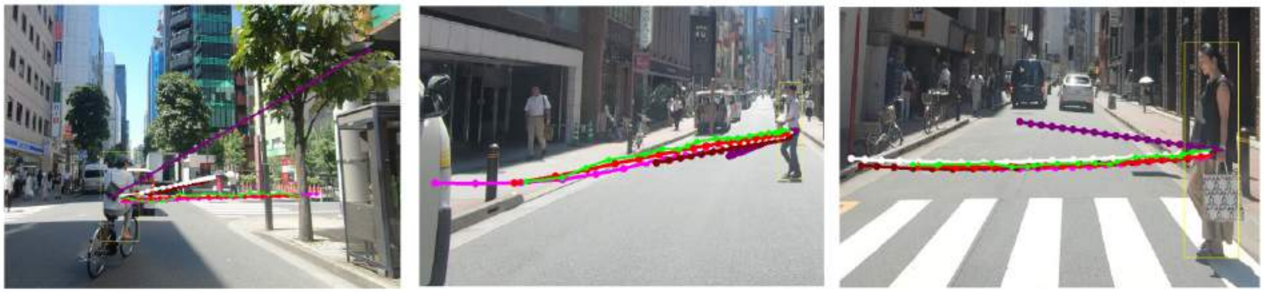


Figure 4: Comparison with others: ground truth ●, Titan_EP+IP+AP (ours) ●, Titan_EP+IP (w/o action) ●, Social-LSTM [1] ●, Social-GAN [2] ○, Const-Vel [3] ●, bounding box at T_{obs} ▭. Images are cropped for better visibility.

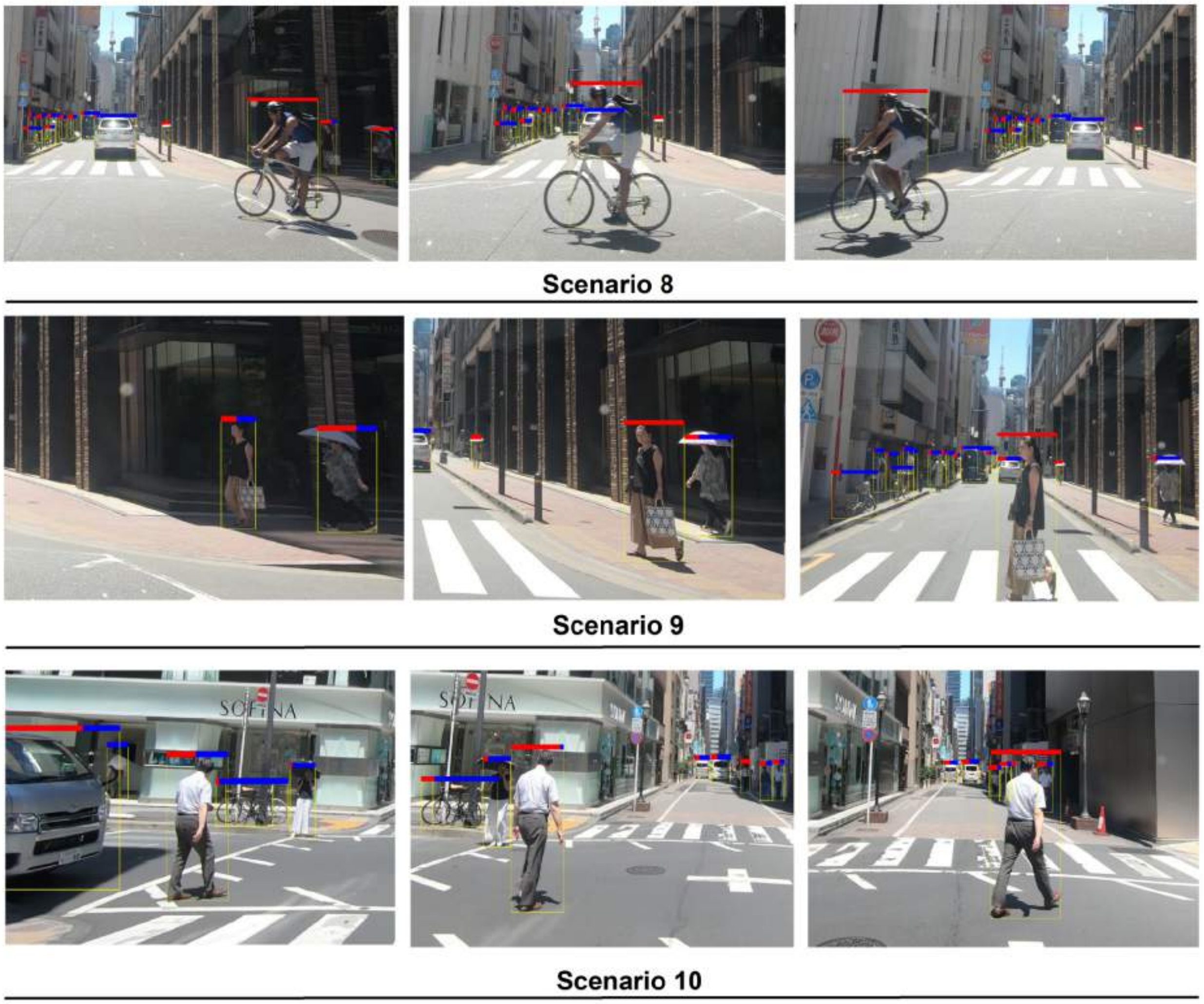


Figure 5: Qualitative results of Importance from different sequences, RED color is high importance, blue is low importance and yellow bounding box is the last observation. (Images cropped for better visibility)

	Method	w/ MT loss↑	w/o MT loss↑
person	atomic	0.9219	0.7552
	simple	0.5318	0.3173
	complex	0.9881	0.9880
	communicative	0.8649	0.8647
	transportive	0.9080	0.9080
	overall	0.8429	0.7667
vehicle	motion	0.9918	0.7130
	trunk	1.00	1.00
	doors	1.00	1.00
	overall	0.9921	0.9043
overall↑		0.8946	0.8127

Table 3: Action recognition results (mAP) on TITAN.

B.3. Additional Qualitative Results

Figure 3 and 4 show the prediction results of the proposed approach for future object localization. Titan_EP+IP+AP consistently shows better performance against the baseline model and the state-of-the-art methods. We also observed that t

In Figure 5, the proposed Agent Importance Module (AIM) is evaluated on additional sequences. The ego-vehicle decelerates due to the crossing agent, and our system considers this agent as having a higher influence (or importance)than other agents. Agents with high importance are depicted with a red over-bar. Particularly in scenario 10, when the person walks along the road in the longitudinal direction, its importance is relatively low. However, the importance immediately increases when the motion changes to the lateral direction.

C. Implementation

TITAN framework is trained on a Tesla V100 GPU using PyTorch Framework. We separately trained action recognition, future object localization, and future ego-motion prediction modules. During training, we used ground-truth data as input to each module. However, during testing, the output results of one module are directly used for later tasks.

C.1. Future Object Localization

During training, we used a learning rate of 0.0001 with RMSProp optimizer and trained for 80 epochs using a batch size of 16. We used hidden state dimension of 512 for both encoder and decoder. A size of 512 is used for the embedding size of action, interaction, ego-motion and bounding box. The input box dimension is 4, action dimension is 8, and ego-motion dimension is 2. The original image size width is 1920 $pixels$ and height is 1200 $pixels$ and accordingly cropped using the bounding box dimension. It is further resized to 228 × 228 for the I3D-based action detector. The bounding box inputs and outputs are normalized between 0 to 1 using image dimensions.

	Layer	Kernal shape	Output shape
0	ego_box_embed.Linear_0	[4, 512]	[1, 10, 512]
1	ego_box_embed.ReLU_1	-	[1, 10, 512]
2	ego_action_embed.Linear_0	[8, 512]	[1, 512]
3	ego_action_embed.ReLU_1	-	[1, 512]
4	ego_motion_embed.Linear_0	[2, 512]	[1, 10, 512]
5	ego_motion_embed.ReLU 1	-	[1, 10, 512]
6	box_encoder.GRUCell_enc	-	[1, 512]
7	motion_encoder.GRUCell_enc	-	[1, 512]
8	int_encoder.embed.Linear_0	[24, 512]	[1, 512]
9	int_encoder.embed.ReLU_1	-	[1, 512]
10	int_encoder.encode.GRUCell_enc	-	[1, 512]
11	concat_to_hidden.Linear_0	[2048, 512]	[1, 512]
12	concat_to_hidden.ReLU_1	-	[1, 512]
13	(8 to 12, repeat based on number of pairwise interactions)		
14	(0 to 12, unroll 10 times)		
15	pred.GRUCell_dec	-	[1, 512]
16	pred.hidden_to_input.Linear_0	[512, 512]	[1, 512]
17	pred.hidden_to_input.ReLU_1	-	[1, 512]
18	pred.hidden_to_output.Linear_0	[512, 10]	[1, 10]
19	pred.hidden_to_output.Sigmoid_1	-	[1, 10]
20	(15 to 19, unroll 20 times)		

Table 4: Future Object Localization model summary with an example batch size of 1

The model summary for Future Object Localization is shown in Table 4. We embed the bounding box (through 0 and 1), action (2-3), ego-motion (4-5) at each time step, and pairwise interaction encoding (8-12). We concatenate the embedded features through (11-12), which are given from the hidden states of the bounding box encoder GRU (6), the hidden states of the ego encoder GRU (7), encoded interaction (10) and action embedding (3). We encode all information for 10 observation time steps from (14). We decode the future locations using decoder GRU for 20 future time steps (20).

C.2. Action Recognition

We used Kinetics-600 pre-trained weights for both I3D and 3D-ResNet. For I3D, we use layers until Mixed_5c layer of the original structure. We used learning rate of 0.0001 and a batch size of 8. We trained it for 100 epochs. The input size is 3 × 10 × 244 × 244, where 10 is the number of time steps, 3 is the number of RGB channels. If the agent is occluded and reappears at any time step, we used the last observed crop of image for that the agent. During training we backpropagate the gradients for pedestrians and vehicles with the loss function as shown below:

$$\mathcal{L}_{total} = \mathbb{1}_p \mathcal{L}_a{}^{i=1:5} + (1 - \mathbb{1}_p) \mathcal{L}_a{}^{i=6:8}, \tag{1}$$

where $\mathbb{1}_p$ is an indicator function that equals 1 if the agent is a pedestrian and 0 if the agent is a vehicle. We refer to the main manuscript for $\mathcal{L}_a$. The model summary for action recognition is shown in Table 5. Note that, from mixed_5c layer [b0, b1b, b2b, b3b] are concatenated to give a shape of [1,1024,2,7,7] which is flattened to give a tensor of shape [1,100352] before feeding it to each MLP head for individual action sets.

	Layer	Kernal shape	Output shape
1	i3d.Conv3d_1a_7x7.conv3d	[3, 64, 7, 7, 7]	[1, 64, 5, 112, 112]
..	.....		
126	i3d.Mixed_5c.b3b.BatchNorm3d	[128]	[1, 128, 2, 7, 7]
127	action.hid_to_pred1.Linear_0	[100352, 10]	[1, 10]
128	action.hid_to_pred1.Softmax_1	-	[1, 10]
129	action.hid_to_pred2.Linear_0	[100352, 13]	[1, 13]
130	action.hid_to_pred2.Softmax_1	-	[1, 13]
131	action.hid_to_pred3.Linear_0	[100352, 7]	[1, 7]
132	action.hid_to_pred3.Softmax_1	-	[1, 7]
133	action.hid_to_pred4.Linear_0	[100352, 4]	[1, 4]
134	action.hid_to_pred4.Softmax_1	-	[1, 4]
135	action.hid_to_pred5.Linear_0	[100352, 4]	[1, 4]
136	action.hid_to_pred5.Softmax_1	-	[1, 4]
137	action.hid_to_pred6.Linear_0	[100352, 4]	[1, 4]
138	action.hid_to_pred6.Softmax_1	-	[1, 4]
139	action.hid_to_pred7.Linear_0	[100352, 3]	[1, 3]
140	action.hid_to_pred7.Softmax_1	-	[1, 3]
141	action.hid_to_pred8.Linear_0	[100352, 3]	[1, 3]
142	action.hid_to_pred8.Softmax_1	-	[1, 3]

Table 5: I3D action recognition model summary with an example batch size of 1

C.3. Future Ego-Motion Prediction

We use batch size of 64, learning rate of 0.0001 and trained for 100 epoch with RMSProp optimizer. We use the hidden state dimension of 128 for both encoder and decoder. We use the embedding size of 128. The prediction is done for 20 time steps in future. The input and output dimensions are 2 at each time step.

The model summary of the future ego-motion prediction is shown in Table 6. We embed the ego motion at each

	Layer	Kernal shape	Output shape
0	ego_embed.Linear_0	[2, 128]	[1, 10, 128]
1	ego_embed.ReLU_1	-	[1, 10, 128]
2	ego_encoder.GRUCell_enc	-	[1, 128]
3	(0 to 3, unroll 10 times)		
4	pred.box_embed.Linear_0	[4, 128]	[1, 1, m, 128]
5	pred.box_embed.ReLU_1	-	[1, 1, m, 128]
6	pred.action_embed.Linear_0	[8, 128]	[1, 1, m, 128]
7	pred.action_embed.ReLU_1	-	[1, 1, m, 128]
8	pred.concat_to_hid2.Linear_0	[256, 128]	[1, 1, m, 128]
9	pred.AIM_layer.Linear_0	[128, 1]	[1, 1, m, 1]
10	pred.AIM_layer.Tanh_1	-	[1, 1, m, 1]
11	pred.concat.concat_0	-	[1, 256]
12	pred.concat_to_hid.Linear_0	[256, 128]	[1, 128]
13	pred.GRUCell_dec	-	[1, 128]
14	pred.hid_to_pred_input.Linear_0	[128, 128]	[1, 128]
15	pred.hid_to_pred_input.ReLU_1	-	[1, 128]
16	pred.Linear_hid_to_pred	[128, 2]	[1, 2]
17	(4 to 15, unroll 20 times)		

Table 6: Future ego motion prediction model summary with an example batch size of 1, m is the number of agents at that future time step

time step (0-1) and use GRU encoder (2) for 10 observation time steps (3). The encoded information is used for the decoder. The embedded future bounding box (4-5) and embedded current action (6-7) are concatenated (8). The agent importance module (AIM) is used to weight the agents at each time step (9-10). We concatenate (11) the AIM output with the past hidden state and embed it (12). The embedded feature is used as an input hidden state. The current hidden state (13) is passed to the next time-step (14-15) using GRU. The output is decoded (16) from the hidden state at each time step (17). As a result, we get for 20 future predictions.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[2] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

[3] Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. The simpler the better: Constant velocity for pedestrian motion prediction. *arXiv preprint arXiv:1903.07933*, 2019.