Supplementary Material
Action-slot: Visual Action-centric Representations for Multi-label Atomic Activity Recognition in Traffic Scenes

Chi-Hsi Kung[1] Shu-Wei Lu[1] Yi-Hsuan Tsai[2] Yi-Ting Chen[1]
National Yang Ming Chiao Tung University[1] Google[2]
{chkung,tomy45651.sc06,ychen}@nycu.edu.tw yhtsai@google.com

Appendix

In this appendix, we cover,

A) Construction of the TACO dataset
B) nuScenes annotation
C) More ablation study of Action-slot
D) Analysis in challenging scenarios
E) Analysis of the TACO Dataset
F) Limitations
G) Implementation details

A. Construction of The TACO Dataset

We introduce the construction of the proposed Traffic Activity Recognition (TACO) dataset. We leverage the CARLA simulator [10] to collect arbitrary traffic scenarios for achieving a balanced activity class distribution in a large scale.

Scenario Collection. Certain atomic activities are rare and difficult to collect in the real world, as shown in the class distribution comparison of TACO and OATS in Figure 2 of the main paper We propose to leverage the CARLA simulator [11] to construct the synthetic dataset. We choose CARLA simulator because it is widely accepted and popular in the computer vision community, where it provides various sensor suites and high-fidelity simulations to facilitate autonomous driving development and test [1, 6, 7, 23, 25, 26, 34, 35], safety-critical scenario generation [9, 16, 24], and domain adaption [28].

Scenarios are collected in the built-in maps (i.e., Town01, Town02, Town03, Town04, Town05, Town06, Town07, and Town10HD) defined in CARLA 0.9.14. We use Town10HD as the testing set. The rest are for the training set. We pinpoint all intersections on each map and subsequently gather specific scenarios related to these intersections. Note that, we also collect T-intersections, following the topology definition discussed in OATS [3], as shown in Figure 1.

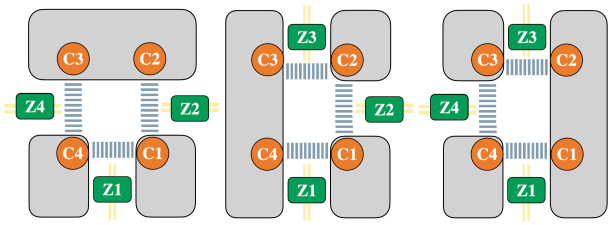


Figure 1. Illustration of the topology in atomic activity for three types of T-intersections.

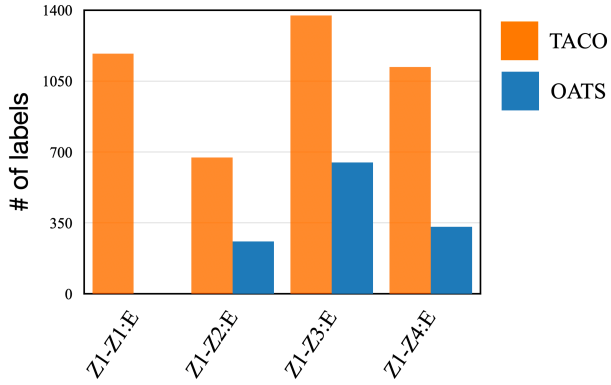


Figure 2. Ego-vehicle's action distribution in TACO and OATS.

We collect scenarios with the following three approaches:

(a) **Auto-pilot** We use the built-in auto-pilot to control ego-vehicle and road users, including vehicles and pedestrians. We do not set destinations for all road users. We program an automatic scenario collection process when (1) ego-vehicle approaches an intersection and (2) ego-vehicle is surrounded by at least one road user.
The length of the recordings varies from 51 to 242 frames. We set the duration for capturing various ego motions. We randomly set the numbers of road users and spawn them on a map. To collect diverse atomic

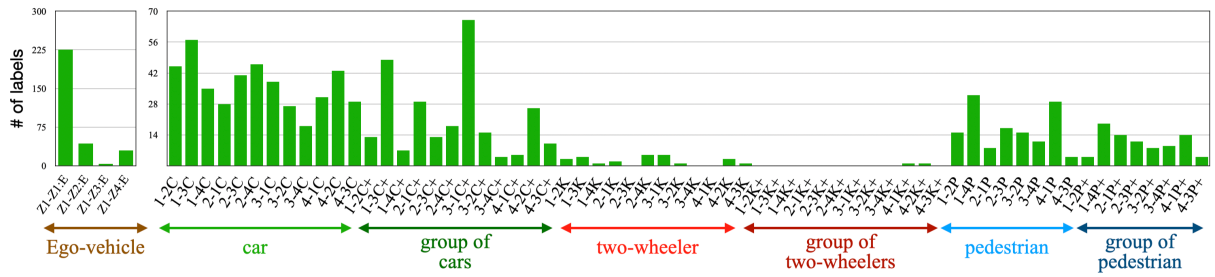


Figure 3. The distribution of atomic activities in the nuScenes dataset [5]. Note that we neglect the notation of topology (i.e., roadway **Z** and corner **C**) in the x-axis due to limited space.

activities, we set all road users (except for the ego-vehicle) to ignore any traffic rules, including traffic lights and stop signs.

(b) **Automatic Scenario Generation [19].** We use existing pre-recorded *basic scenarios* in RiskBench [19] and automatically generate diverse scenarios from a *basic scenario*. Specifically, we select the pre-recorded basic scenarios from *interactive scenarios* where ego-vehicle interacts with other risky road users and *non-interactive scenarios* where ego-vehicle does not interact with any road user. Then we follow the augmentation process proposed in RiskBench [19] to automatically generate diverse scenarios via injecting random road users and changing the weather. The scenarios collected by RiskBench provide more human-like maneuvers and more risky interactions compared to the scenarios collected by auto-pilot.

(c) **Scenario Runner [2].** Scenario runner, a scenario collection tool developed by CARLA [10], can explicitly generate a scripted scenario by defining a set of routes for road users. Although auto-pilot and automatic scenario generation can help collect diverse traffic activities, it is difficult to generate some specific classes of atomic activity frequently. For example, *Z3-Z2:K+*, a group of bicyclists turns left from the opposite roadway of ego-vehicle. Therefore, we leverage scenario runner to explicitly collect the scenarios that are difficult for the auto-pilot method. A scenario starts to collect when ego vehicle reaches a trigger point (i.e., location) predefined in the script. A scenario ends when all scripted actions are accomplished. To further enhance the diversity of scenarios, we randomly spawn road users surrounding the ego vehicle. Note that road users are not guaranteed to be involved in an atomic activity.

For all collection methods, we randomly set the weather and light conditions. Note that we exclude night scenes because of the poor visibility.

Sensor Suites. We deploy a wide field-of-view camera (120 degrees) to record events taking place on the extreme left and right sides of the ego-vehicle. For example, pedestrians crossing the street on the left (*C3-C4:P* and *C4-C3:P*), and a vehicle turns right from the left roadway (*Z4-Z1:C*). We collect the corresponding images and instance segmentation. In addition, we collect instance segmentation from Bird's eye view.

Annotation Criterion. We follow the same annotation criterion defined in OATS [3]. The annotation of atomic activities can be subjective, e.g., whether a turning right car that stops for a crossing pedestrian should be annotated. Thus, the whole annotation work is done by one person to ensure annotation consistency. In addition, we list a set of annotation criteria to enhance the quality of data. If any of the criteria is valid in a video, the annotator is asked to discard the whole scenario.

1. Annotator cannot determine the starting roadway of a road user.
2. Annotator cannot determine a road user's destination.
3. Annotator cannot determine any road user's action in a video.
4. One of the road users has completed an atomic activity at the beginning of the video, e.g., driving away from **Z4** and almost arriving at **Z1**.

We additionally filter out scenarios with zero atomic activity, meaning there are no other road users engaged in any actions. This helps address the positive-negative imbalance issue in multi-label recognition [8, 27, 33]. We also ensure that each atomic activity remains observable for a minimum specified duration. This is due to the fact that the video will be subsampled into a fixed-length K short clip before being inputted into a model. We set K as 16 because the majority of models we benchmarked utilize 16-frame clips. Consequently, for an N-frame video, we ensure that each atomic activity is present for a minimum of N/K frames. Otherwise, annotators will discard the entire scenario.

Annotation for Ego-vehicle's Action. We provide annotations of the ego-vehicle's actions for every scenario. We annotate *Z1-Z1:E* if the annotator is unable to determine the

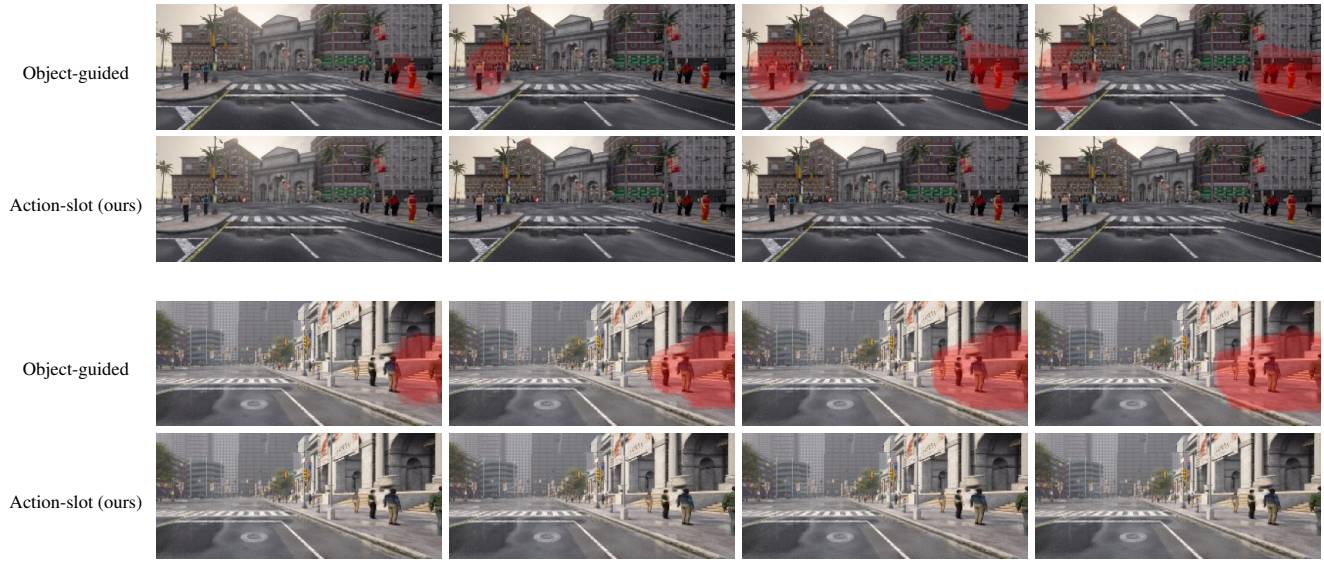


Figure 4. Visualization of attention maps of **Object-guided** and **Action-slot** from slots that predict false positives on additional scenarios. The scenario presents zero atomic activity but many static road users. False positive predicted by the object-guided model: C1-C4:P, C4-C1:P (upper) and C1-C2:P, C2-C1:P, C1-C2:P+ (bottom). The attention scores within the Action-slot do not surpass the threshold in any region.

ego vehicle's action. For instance, when the ego-vehicle remains stationary at a traffic light throughout the entire scenario or moves slowly towards an intersection without exhibiting a specific action. This approach differs from OATS which excludes scenarios lacking a specific ego-vehicle action.

Detailed Dataset Statistics. We have annotated a total of 16,521 instances of atomic activity. The maximum number of road users observed in a single frame is 37. In a video, it reaches up to 63. On average, each video encompasses 2.43 labeled traffic pattern descriptions. The duration of the captured videos varies, ranging from 51 frames to 242 frames. The average length of the videos is 109.341 frames, with a frame rate of 20Hz.

B. nuScenes Annotations

We annotate nuScenes [5], one of the most popular real-world traffic scene datasets, for additional experiments. The dataset was collected in Boston and Singapore. We follow the same annotation criterion in TACO for nuScenes. Specifically, we scan through the *train_val* set of nuScenes and annotate any scenarios in the 4-way and T-intersections. However, the data in nuScenes is collected with a very low frame per second (FPS), we thus select 16 consecutive frames as a short clip. Note that we discard the night scenes because of poor visibility. To this end, we obtain 426 clips and 933 atomic activity labels for the new nuScenes dataset. The atomic activity class distribution is presented in Figure 3. We randomly divide the set with 340 clips for the training set and 86 clips for the testing set. We downsample the image size to 256 $\times$ 768, which is the same as TACO. For transfer learning, we downsample the image size to the same size with pre-trained datasets, i.e., 224 $\times$ 224 for OATS pre-trained and 256 $\times$ 768 for TACO pre-trained. We neglect the atomic activities involved with grouped two-wheelers (K+) when calculating mAP because there are only 2 labels in the whole dataset.

C. More Ablation Study of Action-slot

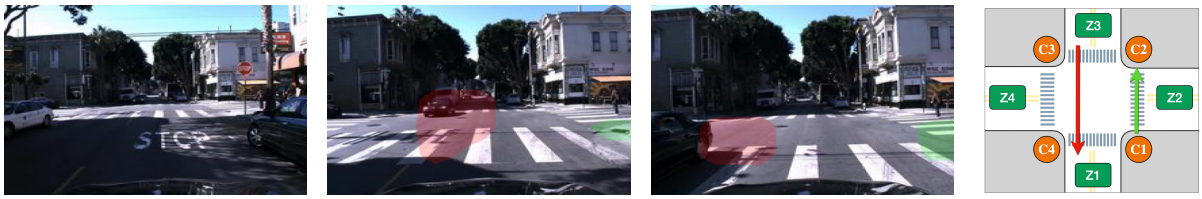
Table 1. Comparison of using cross-attention and slot-attention in Action-slot on OATS. S1, S2, and S3 denote the three test splits in OATS.

	S1	S2	S3	mAP
Cross attention	29.0	32.8	35.4	32.4
Slot attention	**48.1**	**47.7**	**48.8**	**48.2**

Object-guided vs. Action-slot. We further provide detailed insights into the failure cases of object guidance. We hypothesize that the object guidance may mislead the model because not all objects are involved in an activity.

We create scenarios where many pedestrians are static on sidewalks and not involved in any activities, as shown in Figure 4, to better demonstrate the misleading signal caused by the object guidance. We visualize the attention from any

Black vehicles stop at *Z3* in this clip. Atomic activities: C2-C3:P+ and C3-C2:P+.



Many parked cars on both sides of the street. Atomic activities: Z3-Z1:C and C1-C2:P.

Parked car and pedestrians on the left and car stopping temperately at the *Z3*. Atomic activity: Z3-Z2:C.

Figure 5. Action-slot's attention visualization in scenarios where atomic activities and static road users are present. We show scenarios from the TACO, OATS [3], and nuScenes [5] datasets in the first, second, and third row, respectively.

Table 2. Results of Boston and Singapore splits in nuScenes.

	nuScenes	
	Boston	Singapore
X3D [14]	33.7	11.6
ARG [32]	17.2	6.6
Action-slot	**34.7**	**18.3**

slot that predicts **false positive** with red masks. The object-guided method pays attention to the static road users and produces false positive predictions. Moreover, the attention to the false positives is accumulated temporally. We hypothesize this is because the method is not robust to the spatial-temporal features extracted from the backbone [14]. On the other hand, our Action-slot demonstrates the robustness in the scenarios with many static road users.

Cross-Attention vs. Slot Attention. We study the difference between cross-attention and slot-attention for multi-label atomic activity recognition on OATS. Cross-attention can be seen as a query-based method that recently has achieved remarkable success in many tasks [15, 17, 37]. The key difference between cross-attention and slot attention is the dimension to which the softmax operation is applied. The classic cross-attention [31] applies softmax on the tokens, i.e., tokens compete over queries. On the other hand, the softmax in slot attention is applied to the slot dimension, which makes slots compete with each other over the tokens. In Table 1, we conduct experiments by replacing the slot attention in Action-slot with the classic cross-attention. The results show inferior performance of cross-attention compared to Action-slot using slot attention. Moreover, we observe that the background guidance loss L_{bg} in the query-based method can not converge well. The results indicate that the cross-attention method may need a stronger supervision signal.

D. Analysis in Challenging Scenarios

Atomic Activities and Static Road Users in Scenarios. We show qualitative results in scenarios where both atomic activities and multiple static road users are present. In Figure 5, the attention learned by Action-slot focuses on the regions where activities occur instead of being distracted by static road users, e.g., vehicles waiting at traffic lights (first and third row), many parked cars (second row), and pedestrians walking on the sidewalk (third row). The results demonstrate the proposed action-centric representations are robust in crowded traffic scenes and can decompose the atomic activities and non-relevant regions from videos.

Scenario presenting traffic cones in the intersection. Atomic activities: Z4-Z3:C+ and Z1-Z3:K.

Scenario presenting the construction cover the entire corner *C4*. Atomic activities: C4-C1:P.

Figure 6. Action-slot's attention visualization in nuScenes [5]. In the two scenarios, road structures are partially occluded by the traffic cones and construction. Colored masks represented the action slots' attention on distinct activities.

Boston v.s. Singapore. To verify the generalization of Action-slot, we evaluate models on the Boston split and Singapore split in nuScenes [5] in Table 2. Note that the scenarios collected in TACO and OATS [3] are right-hand-traffic. The left-hand-traffic scenarios collected in Singapore thus pose a challenging domain discrepancy for atomic activities. We use the model pretrained on TACO for better performance. Experimental results show that the video-level and object-aware representations both perform inferior in the Singapore split. On the other hand, Action-slot shows the generalization in the scenes with significant domain discrepancy.

Table 3. Comparisons of pretrained representations from OATS and TACO. We perform transfer learning on the nuScenes dataset.

	Kinetics	nuScenes + OATS	+TACO
X3D [14]	19.8	18.9	27.8
ARG [32]	12.2	12.7	17.0
Action-slot	23.6	23.6	**32.3**

Occluded Road Topology. We present the qualitative results to demonstrate that Action-slot can handle challenging scenarios where the road topology is significantly occluded in Figure. 6. We observe Action-slot can accurately predict and localize the activities despite the intersection (upper) and corner (lower) occupied by the traffic cones and construction, respectively. This demonstrates the strong generalization of Action-slot on recognizing road topology and the reasoning ability of road topology.

Road User with Multiple Actions. We find that Action-slot can handle the scenarios where a road user performs multiple actions consecutively. In Figure 7, Action-slot accurately predicts the two atomic activities involved with the crossing pedestrian and spatial-temporally localizes the transition of two actions. The result again demonstrates the effectiveness of the proposed action-centric representations for the action-aware task.

E. Analysis of The TACO Dataset

OATS Pretrain v.s. TACO Pretrain. We compare the pertained representations learned from OATS and TACO by fine-tuning them on nuScenes in Table 3. The results show that models pretrained on TACO outperform the ones pretrained on OATS, verifying the real-world value and transferability of the proposed TACO dataset.

Activity Classes Analysis. We report the performance of models for all 64 classes of atomic activities, which can not be achieved in OATS [3] and nuScenes [5]. We find two interesting observations. First, most results of activities involved with grouped road users (e.g., *C1-C2:C+*) are better than the ones involved with a single road user (e.g., *C1-C2:C*), in which one possible reason is that the larger regions of interest are easier to predict. Second, the smaller or more distant activities are more challenging, which can be attributed to the insufficient representations learned from the backbone.

F. Limitation

Action-slot. We find Action-slot performs less effectively in the occluded scenarios where the activities are visually overlapped. In Figure 8, a bus with *Z2-Z1* action occludes a white car on the right side with action *Z2-Z3*. Action-slot successfully predicts the bus's action *Z2-Z1* but fails to predict *Z2-Z3* and can not localize it via attention. We hypothesize that this is because the occlusion may confuse the competition mechanism in slot attention, i.e., two slots

Figure 7. Action-slot's attention visualization in the TACO scenario where a crossing pedestrian first performs action C4-C3:P then returns with C3-C4:P.

Figure 8. Action-slot's attention visualization in the TACO scenario where an atomic activity is occluded. The yellow bus with red masks partially occludes the white car on the right side (Z2). Action-slot successfully predicts Z3-Z1:C and Z2-Z1:C but misses the occluded white car *Z2-Z3:C* (black arrow in illustration).

Table 4. Results of all 64 classes of atomic activity. Each grouped row is divided by the type of road users involved.

	Z1-Z2:C	Z1-Z3:C	Z1-Z4:C	Z2-Z1:C	Z2-Z3:C	Z2-Z4:C	Z3-Z1:C	Z3-Z2:C	Z3-Z4:C	Z4-Z1:C	Z4-Z2:C	Z4-Z3:C
X3D [14]	21.1	27.1	31.5	28.2	33.4	21.8	28.4	34.5	18.9	22.1	28.8	32.1
ARG [32]	36.2	17.9	23.6	25.9	26.8	14.5	15.3	15.7	16.2	41.4	30.0	19.1
Action-slot	**48.5**	**47.9**	**53.1**	**57.7**	**54.1**	**45.9**	**41.5**	**43.5**	**47.0**	**48.1**	**44.8**	**44.7**

	Z1-Z2:C+	Z1-Z3:C+	Z1-Z4:C+	Z2-Z1:C+	Z2-Z3:C+	Z2-Z4:C+	Z3-Z1:C+	Z3-Z2:C+	Z3-Z4:C+	Z4-Z1:C+	Z4-Z2:C+	Z4-Z3:C+
X3D [14]	76.6	42.7	3.1	61.7	58.7	48.6	41.1	60.9	64.5	76.4	73.4	**70.8**
ARG [32]	32.8	11.0	0.1	38.3	3.8	17.8	9.8	13.2	9.4	4.1	15.6	1.2
Action-slot	**86.9**	**63.7**	**28.7**	**74.9**	**59.5**	**75.2**	**64.6**	**79.0**	**79.1**	**82.1**	**78.0**	69.7

	Z1-Z2:K	Z1-Z3:K	Z1-Z4:K	Z2-Z1:K	Z2-Z3:K	Z2-Z4:K	Z3-Z1:K	Z3-Z2:K	Z3-Z4:K	Z4-Z1:K	Z4-Z2:K	Z4-Z3:K
X3D [14]	17.3	27.9	24.9	36.0	10.6	14.2	30.5	21.5	9.1	13.7	21.7	13.0
ARG [32]	**55.0**	17.3	0.7	50.3	19.2	20.7	22.7	**57.9**	10.3	**46.9**	25.6	23.2
Action-slot	39.9	**54.2**	**45.3**	**51.1**	**30.3**	**41.9**	**46.1**	36.9	**36.9**	35.5	**38.8**	**35.4**

	Z1-Z2:K+	Z1-Z3:K+	Z1-Z4:K+	Z2-Z1:K+	Z2-Z3:K+	Z2-Z4:K+	Z3-Z1:K+	Z3-Z2:K+	Z3-Z4:K+	Z4-Z1:K+	Z4-Z2:K+	Z4-Z3:K+
X3D [14]	76.7	31.0	1.0	**83.5**	26.3	53.2	54.1	**75.0**	36.6	40.4	**79.8**	**55.2**
ARG [32]	22.1	**22.1**	20.6	5.0	11.3	3.9	24.2	7.0	2.5	5.3	20.4	9.1
Action-slot	**93.7**	**56.7**	7.7	67.3	**51.9**	**65.7**	**74.3**	74.0	**68.3**	**60.6**	76.8	48.4

	C1-C2:P	C1-C4:P	C2-C1:P	C2-C3:P	C3-C2:P	C3-C4:P	C4-C1:P	C4-C3:P
X3D [14]	34.4	43.5	38.9	35.0	25.1	29.6	45.5	27.6
ARG [32]	31.1	38.7	23.7	12.8	10.6	22.5	39.3	15.3
Action-slot	**52.4**	**60.6**	**55.9**	**42.5**	**44.3**	**38.2**	**65.2**	**34.2**

	C1-C2:P+	C1-C4:P+	C2-C1:P+	C2-C3:P+	C3-C2:P+	C3-C4:P+	C4-C1:P+	C4-C3:P+
X3D [14]	39.2	53.3	47.3	24.2	23.2	32.4	62.6	29.1
ARG [32]	19.6	24.8	18.0	6.1	5.7	9.4	24.6	8.4
Action-slot	**61.0**	**74.5**	**65.0**	**34.8**	**33.0**	**37.4**	**80.8**	**35.3**

compete over the overlapped regions. We hope our findings can inspire the community to discover more advanced action-centric representations that can handle occlusion issues.

The TACO Dataset. We observe the 64 classes of atomic activities in TACO can not fully cover the diverse events in traffic scenes. For example, vehicles can only move between roadways and pedestrians can only move between two near corners. However, two-wheelers can move between corners, e.g., *C1-C2:K+*, and pedestrians can also move diagonally, e.g., *C1-C3:P*. These atomic activities are important to many applications, such as safety-critical scenario generation [16, 24, 34]. However, the existing autopilot mechanism in the CARLA simulator does not support the collection of such atomic activities. We aspire for our research to inspire collaborative efforts within the community to improve existing atomic activity datasets. This involves gathering larger, more diverse datasets from real-world scenarios and advancing the sophistication of simulator-based auto-pilots.

G. Implementation Details

Data Preprocessing. We downsample a video into a fixed-length short clip as models' input, which is common practice in video recognition [3, 22, 27, 29]. Specifically, we randomly sample subsequent with uniform intervals between frames for the training set and fix the subsequent for the testing set. Note that we neglect this process for nuScenes because we annotate each sample as a fixed 16-frame clip.

Ego-vehicle's action. To enhance the awareness of ego motion, we include a module for all models to predict ego-vehicle's action via global features. The global features $F_{ego} \in \mathbb{R}^{256}$ are generated by applying a Conv3D with kernel size 1. Since each video must have a label for the corresponding ego-vehicle's action, the models output a multi-class prediction with a fully connected layer and softmax operation. Note that we neglect the ego vehicle's action prediction in OATS [3] since the absence of the annotation in the released dataset. In this paper, we do not report the accuracy of predicting ego-vehicle's pattern because the performance of all models saturates with nearly 100%.

Architecture of Action-slot. In this work, we eliminate the GRU from the slot updating process [4, 12, 18, 20, 36] for Action-Slot. This is because of its negligible impact on enhancing performance in our experiments.

Training for transformer-based methods. We freeze the pretrained transformer blocks except for the first three and last three blocks during training. We downsample the input image to 224×224 for both MViT [13] and VideoMAE [30] on all datasets. We attempt to adapt the pre-trained positional embedding of MViT as suggested in ViT [11] by interpolating spatial positional embedding. For example, we adapt token numbers from 7×7 to 8×24 in our TACO dataset. However, we find the performance degraded with interpolation. We thus simply use image size 224×224 to match the original positional embedding size.

Training for object-aware methods. We follow OATS [3] to set the number of object proposals to 20. Specifically, we select the 20 largest bounding boxes in each frame. To match the ground-truth atomic activity labels with the length of proposals, we pad the ground truth with the negative class and use a Hungarian matcher to associate them during training. It is worth noting that because object-aware models output multi-class results for each proposal, we rearrange the outputs to a set for calculating the metrics.

Training for slot-based methods. In order to adapt slot-based baselines to atomic activity recognition, we use the last state of slots as the input to the classifier for the recurrent fashion, i.e., SAVi [12, 18] and MO [4]. As for Slot-VPS [36], we sum up the slots across temporal dimensions. For all slot-based baselines, including our Action-Slot, we set both the dimensions of slots and image features to 256.

Backbone modification for Action-Slot. We use the features of the last convolution block as the input to action slots for all backbone encoders except for SlowFast. SlowFast processes two paths: path *Fast* takes an original input sequence (16 frames) as input and the path *Slow* takes the subsequence with $1/4$ length (4 frames). We apply a pooling operation to the output of path *Fast* for aligning the length of path *Slow* and combine them with channel-wise concatenation. We freeze the entire backbone except for the last ConvBlock.

Hyperparameters. All the models including our Action-Slot are trained for 50, 100, and 100 epochs on OATS [3], TACO, and nuscenes [5], respectively. We use AdamW optimizer [21] with a batch size of 8. The learning rate varies from 1e-4 to 5e-5 and weight decay varies from 0.1 to 0.0001. We apply a 50% dropout to all models' last features layer, e.g., X3D's last ConvBlock. We conduct all experiments with a single NVIDIA 3090 GPU with 24GB.

References

[1] CARLA Autonomous Driving Challenge. https://carlachallenge.org/, 2022. 1

[2] ScenarioRunner for CARLA. https://github.com/carla-simulator/scenario_runner, 2023. 2

[3] Nakul Agarwal and Yi-Ting Chen. Ordered atomic activity for fine-grained interactive traffic scenario understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8624–8636, 2023. 1, 2, 4, 5, 7

[4] Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Discovering Objects that Can Move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11789–11798, 2022. 7

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 4, 5, 7

[6] Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17222–17231, 2022. 1

[7] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. Neat: Neural attention fields for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15793–15803, 2021. 1

[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 2

[9] Wenhao Ding, Haohong Lin, Bo Li, and Ding Zhao. Causalaf: causal autoregressive flow for safety-critical driving scenario generation. In *Conference on Robot Learning*, pages 812–823. PMLR, 2023. 1

[10] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 1, 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 7

[12] Gamaleldin Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C Mozer, and Thomas Kipf. SAVi++: Towards End-to-End Object-centric Learning from Real-world Videos. *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 7

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6824–6835, 2021. 7

[14] Christoph Feichtenhofer. X3D: Expanding Architectures for Efficient Video Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–213, 2020. 4, 5, 6

[15] Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5496–5506, 2023. 4

[16] Niklas Hanselmann, Katrin Renz, Kashyap Chitta, Apratim Bhattacharyya, and Andreas Geiger. King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In *European Conference on Computer Vision*, pages 335–352. Springer, 2022. 1, 7

[17] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 4

[18] Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-Centric Learning from Video. In *International Conference on Learning Representations (ICLR)*, 2022. 7

[19] Chi-Hsi Kung, Chieh-Chi Yang, Pang-Yuan Pao, Shu-Wei Lu, Pin-Lun Chen, Hsin-Cheng Lu, and Yi-Ting Chen. Riskbench: A scenario-based benchmark for risk identification. *arXiv preprint arXiv:2312.01659*, 2023. 2

[20] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric Learning with Slot Attention. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:11525–11538, 2020. 7

[21] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. 7

[22] Tushar Nagarajan, Santhosh Kumar Ramakrishnan, Ruta Desai, James Hillis, and Kristen Grauman. Egoenv: Human-centric environment representations from egocentric video. *arXiv preprint arXiv:2207.11365*, 2022. 7

[23] Eshed Ohn-Bar, Aditya Prakash, Aseem Behl, Kashyap Chitta, and Andreas Geiger. Learning situational driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[24] Davis Rempe, Jonah Philion, Leonidas J. Guibas, Sanja Fidler, and Or Litany. Generating useful accident-prone driving scenarios via a learned traffic prior. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 7

[25] Katrin Renz, Kashyap Chitta, Otniel-Bogdan Mercea, A Koepke, Zeynep Akata, and Andreas Geiger. Plant: Explainable planning transformers via object-level representations. *arXiv preprint arXiv:2210.14222*, 2022. 1

[26] Hao Shao, Letian Wang, Ruobing Chen, Steven L Waslander, Hongsheng Li, and Yu Liu. Reasonnet: End-to-end driving with temporal and global reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13723–13733, 2023. 1

[27] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 510–526. Springer, 2016. 2, 7

[28] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. Shift: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21371–21382, 2022. 1

[29] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. *arXiv preprint arXiv:2301.02217*, 2023. 7

[30] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022. 7

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Conference on Neural Information Processing Systems*, 30, 2017. 4

[32] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning Actor Relation Graphs for Group Activity Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4, 5, 6

[33] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[34] Chejian Xu, Wenhao Ding, Weijie Lyu, Zuxin Liu, Shuai Wang, Yihan He, Hanjiang Hu, Ding Zhao, and Bo Li. Safebench: A benchmarking platform for safety evaluation of autonomous vehicles. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1, 7

[35] Jimuyang Zhang, Zanming Huang, and Eshed Ohn-Bar. Coaching a teachable student. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7805–7815, 2023. 1

[36] Yi Zhou, Hui Zhang, Hana Lee, Shuyang Sun, Pingjun Li, Yangguang Zhu, ByungIn Yoo, Xiaojuan Qi, and Jae-Joon Han. Slot-VPS: Object-centric Representation Learning for Video Panoptic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3093–3103, 2022. 7

[37] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17863–17873, 2023. 4