Supplementary Material for Exploring Data Aggregation in Policy Learning for Vision-based Urban Autonomous Driving

Aditya Prakash[1] Aseem Behl[*1,2] Eshed Ohn-Bar[*1,3] Kashyap Chitta[1,2] Andreas Geiger[1,2]
[1]Max Planck Institute for Intelligent Systems, Tübingen [2]University of Tübingen [3]Boston University
{firstname.lastname}@tue.mpg.de

Abstract

In this supplementary document, we first study the common failure cases of behavior cloning in dense urban scenarios. We then provide implementation details of the architecture and the training procedure used in the main paper. Further, we present a theoretical analysis of the proposed approach and provide additional results paired with a more in-depth analysis of the findings from the main paper. Finally, we bring to light several limitations in the NoCrash benchmark and evaluation protocol for the CARLA simulator. The supplementary video contains qualitative comparisons of our approach (DA-RB$^+$) against CILRS$^+$ in different weather conditions and traffic scenarios.

1. Failure Cases of Behavior Cloning

In this section, we study the common failure cases of behavior cloning in dense urban scenarios. For this, we consider the conditional imitation learning model [4] which is the current state-of-the-art on the CARLA 0.8.4 NoCrash benchmark. The common failure cases correspond to collision with pedestrians, collision with vehicles and traffic light violations. In most of these scenarios, we observe that the driving policy is not able to brake adequately as shown in the Fig. 1. We also provide driving videos of these scenarios in the attached video.

2. Implementation Details

In this section, we first give a detailed description of the architecture used in our approach. We then describe the loss function and the training protocol employed in our approach.

2.1. Architecture

We build on the conditional imitation learning framework of [4] and use the exact same architecture (Table 1) as that of CILRS [4] model. The input to the model consists of an image of resolution 200x88 and the current speed measurement. The image is processed by a ResNet34-based perception module resulting in a latent embedding of 512 dimension. The speed input is processed by two fully-connected layers of 128 units each and combined with the ResNet output using another fully-connected layer of 512 units. This joint embedding is then passed as input to the command branches and the speed branch which output the control values and the predicted speed, respectively. Each of these branches consists of two fully-connected layers with 256 units. We apply a dropout of 0.5 to the last fully-connected layer in each of the branches. CARLA [5] also provides access to four high level navigational commands - (i) turn left, (ii) turn right, (iii) go straight (at intersection) and (iv) follow lane. These high level commands are used as input to a conditional module which selects one of the four command branches to output the control, which consists of steer, throttle and brake.

2.2. Loss Function

The network is trained in a supervised manner with the loss function consisting of two components - (i) **Imitation Loss**: To imitate the expert actions, we use the L_1 loss between the predicted control $\pi(s)$ and the expert control $\pi^*(s)$. This

[*] indicates equal contribution, listed in alphabetical order

Figure 1: **Common failure cases of behavior cloning in urban environments.** Left: collision with pedestrians. Middle: collision with vehicles. Right: traffic light violation. Note the major deviation in brake values compared to expert.

Module	Input	Output
Perception	ResNet34 [7]	512
Measured Speed	1	128
	128	128
	128	128
Joint Input	512 + 128	512
Command branch	512	256
	256	256
	256	3
Speed Prediction	512	256
	256	256
	256	1

Table 1: **Conditional Imitation Learning Architecture [4].**

is represented as $\mathcal{L}_{imitation} = \|\pi(s) - \pi^*(s)\|_1$. (ii) **Speed Loss**: Expert demonstrations have an inherent inertia bias, where most of the samples with low speed also have low acceleration. It is critical to not overly correlate these since the vehicle would prefer to never start after slowing down. This issue can be alleviated by predicting the current vehicle speed as auxiliary task [4]. Therefore, we also use a speed prediction loss, given by $\mathcal{L}_{speed} = \|v - \hat{v}\|_1$ where v is the actual speed, $\hat{v}$ is the predicted speed and $\|\cdot\|_1$ denotes the L_1 norm. The final loss function is a weighted sum of the two components, with a scalar weight λ, given by $\mathcal{L} = \mathcal{L}_{imitation} + \lambda \cdot \mathcal{L}_{speed}$. Following [4], we set $\lambda = 0.008$.

2.3. Data Generation

We use the standard CARLA 0.8.4 data-collector framework[1] for generating data. We consider 4 weather conditions - 'ClearNoon', 'WetNoon', 'HardRainNoon' and 'ClearSunset' - for generating a total of 10 hours of expert training data and 2 hours of validation data in 'Town01' setting with the number of vehicles in the range [30, 60] and number of pedestrians in the range [50, 100]. The expert policy used in the data generation process consists of an A* planner followed by a

[1] https://github.com/carla-simulator/data-collector

PID controller and is provided by the official data collector. The images are rendered at a resolution of 800x600, and then processed to a resolution of 200x88 as in [4].

2.4. Training Protocol

We use the conditional imitation learning framework[2] provided by the authors of [4] for training all methods mentioned in the paper. In all experiments, we use the Adam [9] optimizer and the exact same hyper-parameters as in the original CILRS [4] model. We save the model checkpoints after every 10000 iterations and stop training once the validation loss has stopped improving for 5 consecutive checkpoints. For all iterative algorithms mentioned in the paper, we initialize the behavior policy in each iteration with the trained policy of the previous iteration.

3. Theoretical Analysis

3.1. Performance Guarantees

DAgger [14] is known to have a better performance bound (Eq. (1)) on the total cost incurred over the time horizon compared to behavior cloning (Eq. (2)), which is given by

$$J(\pi) \leq J(\pi^*) + uT\epsilon_N + O(1)\,, \quad \epsilon_N = \min_{\pi \in \Pi} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{s \sim d_{\pi_i}} \ell(s,a) \tag{1}$$

$$J(\pi) \leq J(\pi^*) + T^2\epsilon\,, \quad \epsilon = \mathbb{E}_{s \sim d_{\pi^*}} \ell(s,a) \tag{2}$$

where $J(\pi) = \sum_{i=0}^{T-1} \mathbb{E}_{s \sim d_\pi} [\mathbb{E}_{a \sim \pi(s)} \ell(s,a)]$ is the total cost incurred by the policy π over the time horizon T, $\ell(s,a)$ is a convex upper bound on the (in general non-convex) loss function $\tilde{\ell}(s,a)$ and u upper bounds $Q_t^{\pi^*}(s,a) - Q_t^{\pi^*}(s,\pi^*(s))$ for all $a \in \mathcal{A}$, $s \in \mathcal{S}$ and $t \in \{0, ..., T-1\}$.

However, as described in [2], u in Eq. (1) may be $O(T)$, e.g., if there are *critical states* s such that failing to take the action $\pi^*(s)$ in s results in forfeiting all subsequent rewards. For example, in dense urban driving, these critical states correspond to scenarios involving close proximity to pedestrians and vehicles resulting in collision and termination of episode, so $u = O(T)$. In the presence of these type of scenarios, DAgger has a bound of $O(T^2)$ (Eq. (1)) which is the same as that of behavior cloning. Moreover, this bound can be improved to $O(T)$ (Eq. (3), see [2] for more details) by performing accurately on the critical states.

$$J(\pi) \leq J(\pi^*) + T\epsilon_N + O(1)\,, \quad \epsilon_N = \min_{\pi \in \Pi} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{s \sim d_{\pi_i}} \ell(s,a) \tag{3}$$

The DA-CS variant of our approach explicitly samples the critical states during the aggregation process, thereby increasing the proportion of these states in the training data distribution leading to policies that perform better in these difficult scenarios. Therefore, assuming a convex upper bound on the loss function, DA-CS has better performance guarantees on the total cost incurred over the time horizon compared to DAgger.

While adaptive sampling methods in a mixture of distributions setting are known to have convergence guarantees [3], the theoretical analysis of the performance guarantees for the DA-RB variant of our approach is beyond the scope of this work.

3.2. Mixture of Distributions with Adaptive Sampling

Consider a mixture of k sampling distributions $p_1, ..., p_k \in \Delta_n$ in which the probability of sampling a particular state s is given by $w^T p(s)$, where $w \in \Delta_k$ is the mixture weight vector, Δ_n is the n-dimensional probability simplex, $w(s) := [w_1(s), ..., w_k(s)]$ and $p(s) := [p_1(s), ..., p_k(s)]$. In online policy learning setting, $p_1, ..., p_k$ correspond to the distribution induced by the driving policy in each online iteration. Therefore, in the case of our autonomous driving application, the mixture distribution can be represented by $p_{\pi^*}, p_{\pi_1}, ..., p_{\pi_k}$ where π^* is the expert policy and π_i is the driving policy trained in the i^{th} iteration. In this regard, DAgger [14] and SMILe [13] can also be interpreted in terms of mixture of distributions. While the weight vector $w(s)$ is chosen arbitrarily in the former, $w(s)$ is defined by Eq. (4) in the latter, where β is selected as described in [13].

$$w_{\pi^*}(s) = (1-\beta)^k\,, \quad w_{\pi_i}(s) = \beta(1-\beta)^{i-1} \;\forall\; i \in 1, ..., k \tag{4}$$

However, the weight vectors in DAgger and SMILe are independent of the learned policies which leads to redundancy in the sampled states resulting in non-optimal performance. To rectify this problem, it is desirable to define the weight vector as a

[2] https://github.com/felipecode/coiltraine

Weather	CILRS$^+$	DA-RB$^+$(E)	CILRS$^+$	DA-RB$^+$(E)
	Training Conditions		New Town	
ClearNoon	15	**17**	7	**9**
WetNoon	12	**20**	**4**	**4**
HardRainNoon	9	**17**	6	**14**
ClearSunset	10	**18**	5	**13**
	New Weather		New Town & Weather	
CloudyNoon	**15**	**15**	10	**17**
WetCloudyNoon	10	**16**	8	**15**
MidRainNoon	9	**12**	5	**12**
SoftRainNoon	**14**	13	4	**10**
CloudySunset	13	**17**	10	**15**
WetSunset	8	**15**	**7**	3
WetCloudySunset	9	**14**	11	**14**
MidRainSunset	3	**14**	0	0
HardRainSunset	**4**	**4**	0	0
SoftRainSunset	11	**16**	**9**	5

Table 2: **Performance comparison of DA-RB$^+$(E) and CILRS$^+$ for different weathers.** We report the number of successful episodes (out of 25) on all weathers in the dense setting of all evaluation conditions.

function of the learned policy which results in adaptive sampling of on-policy states in each iteration. Our proposed approach with critical states and replay buffer is one specific instance of adaptive sampling in a mixture of distribution setting. We propose to sample critical states from the on-policy data which is given by

$$\mathcal{S}_c = \left\{ s_c \in \mathcal{S} \,\middle|\, H(s_c, \pi, \pi^*) > \alpha \cdot \max_s H(s, \pi, \pi^*) \right\} \tag{5}$$

where $\mathcal{S} = \{s \,|\, s \sim P(s|\pi)\}$ is the set of states sampled from the distribution $P(s|\pi)$, $H(s, \pi, \pi^*)$ is the sampling criterion and $\alpha < 1$ is chosen empirically. The mixture weight vector is defined as $w_\pi(s) = f(s, \alpha, \pi, H)\ \forall\ \pi \ \in \{\pi^*, \pi_1, ..., \pi_k\}$ where $f(\cdot)$ is implemented using critical states and replay buffer mechanisms as described in Section 3 and Algorithm 1 of the main paper. A natural extension of our approach is to learn the mixture weight vector $w(s)$ in order to optimize for the driving performance. This constitutes a dual optimization problem where the driving policy and the weight vector $w(s)$ are learned in an alternating fashion. Based on the recent works in data distribution optimization [8, 15], we expect this to be a promising direction for future research.

4. Additional Experimental Results

4.1. Weather-wise Performance Breakdown

We provide the performance breakdown of DA-RB$^+$(E) over individual weather conditions (Fig. 2) and compare against the CILRS$^+$ baseline. The evaluation consists of 25 episodes for each weather condition with different start locations and destinations. The results in Table 2 show that our approach outperforms CILRS$^+$ [4] in most of the conditions. However, *MidRainSunset* and *HardRainSunset* weather conditions are especially challenging for behavior cloning since none of the approaches are able to complete even a single episode out of 25. This is due to the presence of extreme conditions, such as excessive shadows, reflections of the buildings in water puddles and glare from sunset which severely complicates perception.

4.2. Considering Traffic Light Violation as Failure Case

In the NoCrash [4] benchmark traffic light violation is not considered as a termination scenario. However, traffic light violation can lead to fatal accidents especially in dense urban setting due to the presence of high number of pedestrians & vehicles. Therefore, obeying traffic lights is an essential part of urban driving which needs to be learned by the driving policy. In this experiment, we consider traffic light violation as a failure case and compare against CILRS$^+$ model. We report the results in Fig. 3 on the dense setting on all the evaluation conditions. Our experiments show that our approach enables the policy to better detect traffic lights.

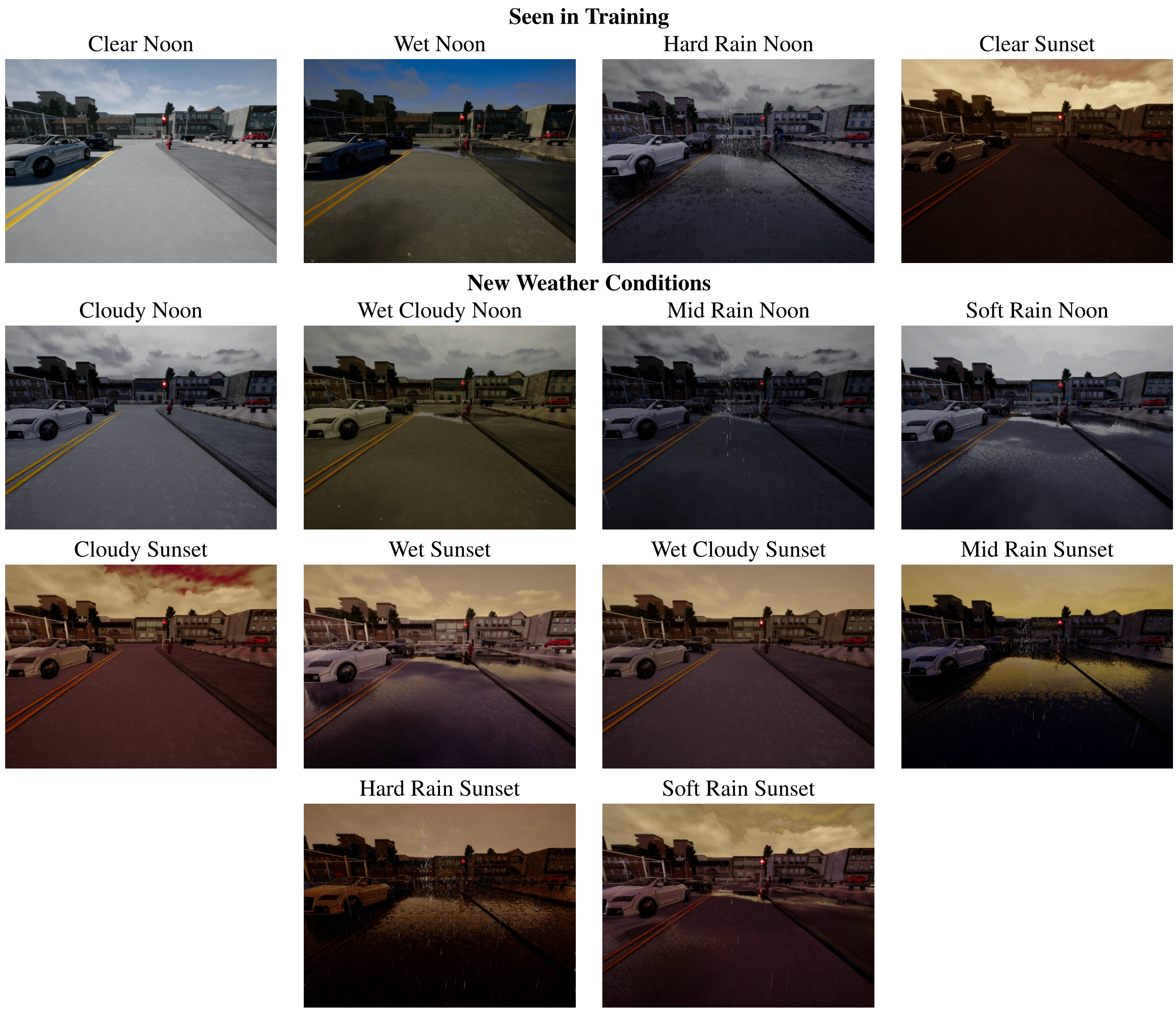


Figure 2: **Visualization of all weather conditions in CARLA 0.8.4.**

4.3. Comparison of SMILe against DAgger

We implement DAgger as per Algorithm 3.1 of [14]. In each iteration of DAgger, we append 2 hours of on-policy data to the current iteration dataset. For SMILe, we follow Algorithm 4.1 of [13] with $\alpha = 0.2$. We execute both algorithms using the same initialization for fair comparison. In our experiments (Fig. 3 in the main paper), we observe that the performance of SMILe is either as good as DAgger in New Weather and New Town & Weather conditions or slightly better in Training and New Town conditions. This is in contrast to the results in [14] where the authors show DAgger to be empirically superior to SMILe. We have shown in the main paper that DAgger is not effective for dense urban driving since the aggregation process does not address the dataset bias issue. However, the training dataset in each iteration in SMILe is sampled from a mixture of policies which leads to better diversity compared to direct aggregation. Also, we observe that SMILe^{+} generalizes very well to New Town and New Town & Weather conditions. This happens due to 2 reasons, (1) Triangular perturbations contribute to the diversity of the data since they simulate off-road drift which is seldom present in the expert's state distribution, (2) SMILe returns an ensemble of policies trained in each iteration which leads to increased robustness and better generalization.

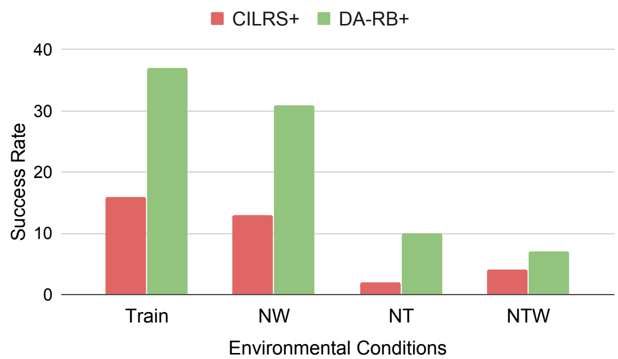


Figure 3: **Success rate when considering traffic light violation as a failure case.** NW - New Weather, NT - New Town, NTW - New Town & Weather.

4.4. Comparison of DART against Triangular Perturbations

For implementing DART [11], we closely follow the code provided by the authors of [11][3]. The performance of DART is quite similar to that of CILRS$^+$ in most of the evaluation conditions (Fig. 3 and Table 3 in the main paper). DART uses a noise model which is optimized to iteratively minimize the covariate shift. These perturbations manifest most prominently in the steering of the vehicle, thereby simulating off-road drift. This is identical to the behavior modeled by triangular perturbations in the steering, therefore, leading to similar results.

4.5. Data Distribution Statistics for Different Sampling Methods

We report statistics regarding the data distribution induced by the sampling mechanisms (Section 3.3 and 4.4 in the main paper) to provide insights into the different type of scenarios captured by the sampling strategies. We focus on two type of statistics - (i) weather-wise data distribution over high level navigational commands (Fig. 6): We report the number of images in the training data for 'follow lane', 'turn left', 'turn right' and 'go straight (at intersection)' navigational commands, (ii) weather-wise data distribution over control values (Fig. 7): We bin the control values into 4 categories - 'brake', 'steer left', 'go straight' and 'steer right'. For the 'brake' category we consider the states where brake is emphasized by the expert policy. Even though brake can have a continuous value in the range [0, 1], we observe that the brake distribution is highly skewed towards the extreme values. Moreover, while visualizing the driving performance, we notice that even a small value produces substantial braking effect. Therefore, we bin all the states where the brake $>$ 0.1 in the 'brake' category. The other 3 categories are defined based on the steering values, which belong in the range [-1, 1]. We bin all the states where steer $<$ -0.1 into 'steer left', steer $\in$ [-0.1, 0.1] into 'go straight' and steer $>$ 0.1 into 'steer right' categories. We prioritize the 'brake' category over the steering categories since braking is the most crucial action to avoid collisions and other failure cases.

From the results in Table 5 of the main paper and Fig. 6, we can observe that the weather 'ClearSunset' and the navigational command 'Go straight (at intersection)' are most correlated with the generalization performance since the sampling method based on absolute error on brake (AE_b) results in the best generalization performance of the driving policy in terms of success rate. This is also apparent by the results of uncertainty-based sampling since it generalization performance is inferior compared to other sampling approaches. Furthermore, from Fig. 7, we can see that a uniform distribution over the control categories and training weathers results in the best generalization performance. Fig. 7 also provides additional insights into the inferior generalization performance of the uncertainty-based sampling approach. Even though uncertainty-based sampling is effective in capturing states where the brake is emphasized, its distribution is highly skewed. This results in the driving policy being overly cautious and braking excessively due to which the policy times out frequently and is not able to successfully complete the episode. This is in line with existing findings in literature on uncertainty-based sampling that highlight the benefits of additionally incorporating diversity among samples as a criteria for data selection [6, 10].

[3]https://github.com/BerkeleyAutomation/DART

4.6. Data Distribution Statistics for DAgger$^+$ and DA-RB$^+$

In this experiment, we provide a qualitative comparison of the different types of scenarios captured by DAgger$^+$ and DA-RB$^+$ to gain more insights into the effectiveness of DA-RB$^+$ against DAgger$^+$. We report statistics regarding the data distribution induced by DAgger$^+$ and DA-RB$^+$ according to the criteria described in Section 4.5. From Fig. 8 and Fig. 9, we observe that the data distribution induced by DAgger$^+$ is very similar to the 10 hours of expert perturbed data used for training CILRS$^+$. This provides further justification for the comparable performance of DAgger$^+$ and CILRS$^+$ (in terms of success rate, failure cases and variance due to training seeds) and corroborates our claim that DAgger is not optimal for urban driving. In contrast, DA-RB$^+$ leads to a uniform distribution over controls (Fig. 9) and emphasizes the sampling of states from the weather 'ClearSunset' and the navigational command 'Go straight (at intersection)' (Fig. 8). Moreover, DA-RB$^+$ is able to capture the critical states where brake is emphasized resulting in trained policies which drive cautiously. This highlights the importance of critical states and uniform data distribution for improved driving in dense urban scenarios.

4.7. Weather-wise Breakdown of Variance Results

In Section 4.3 and Table 4 of the main paper, we report mean and standard deviation of success rate wrt. 5 random training seeds on the dense setting of New Town & Weather. Here, we provide a weather-wise breakdown of the variance results for success rate as well as the failure cases to gain further insights into the robustness of the driving policy in individual weather conditions. From Fig. 10, we observe that DA-RB$^+$ outperforms DAgger$^+$ and CILRS$^+$ on all weather conditions in terms of success rate & collision metrics and leads to cautious driving policies. Moreover, DA-RB$^+$ results in stable training as evidenced by lower standard deviation in all weather conditions (Fig. 11). In contrast, DAgger$^+$ exhibits higher standard deviation compared to CILRS$^+$ on all weather conditions (Fig. 11), especially in collision scenarios. This trend is not visible in Table 4 of the main paper since the values are averaged over 10 weathers. However, when analyzing individual weather conditions, it is apparent that DAgger is not effective in reducing variance and leads to unstable policies. This further exacerbates the limitations of DAgger for dense urban driving. While DA-RB$^+$ is able to reduce variance on success rate and collision metrics, we observe a slight increase in standard deviation of timed out scenarios in some weather conditions. This is due to the presence of exogenous factors such as non-optimal and non-deterministic behavior of dynamic agents which severely influences timed out cases, e.g., multiple vehicles clogging the lane resulting in no space available for driving.

4.8. GradCAM Attention Maps

We examine the GradCAM [16] attention maps of DA-RB$^+$ qualitatively to visualize the region in the image which is important for vehicle control and compare against CILRS$^+$. Specifically, we backpropagate the gradients from the brake signal since it is very important for preventing collisions. The attention maps (Fig. 13) show that our approach enables the driving policy to focus more on the essential aspects of the scene, thereby learning a better implicit visual representation of the environment for urban driving.

5. Limitations of the NoCrash Benchmark

In this section, we highlight the limitations of the NoCrash benchmark [4]. Specifically, we focus on 4 issues - bias against cautious policies, performance metrics, distribution of weathers in the evaluation setting and variance due to training seeds.

5.1. Bias against Cautious Driving Policies during Evaluation

The NoCrash benchmark[4] defines success in terms of the ability of the driving policy to complete the episode within the specified time limit which is computed as the time required to traverse the shortest path from the starting location to the destination at a constant speed of 5 km/hr. However, this time limit does not take into account the presence of dynamic agents and traffic constraints which can severely affect the ability of the driving policy to adhere to the time limit. In particular, we observe that a driving policy which follows all traffic regulations (i) stops for 5-8 seconds on average in case of a red light which significantly increases the probability of getting timed out, (ii) gets obstructed multiple times due to the presence of high density of pedestrians and vehicles to avoid collisions, (iii) is unable to move due to exogenous factors arising from non-optimal and non-deterministic behavior of dynamic agents, e.g., multiple vehicles clogging the lane resulting in no space available for driving. These factors significantly reduce the success rate giving a false impression that the driving policy is failing regularly, when in fact the driving policy is not at fault in these scenarios.

To highlight this problem, we run an evaluation in which we allow the expert policy to ignore the traffic lights (Expert no TL) and compare against an expert policy which follows traffic regulations (Expert). The expert policy consists of an A*

[4] https://github.com/felipecode/coiltraine

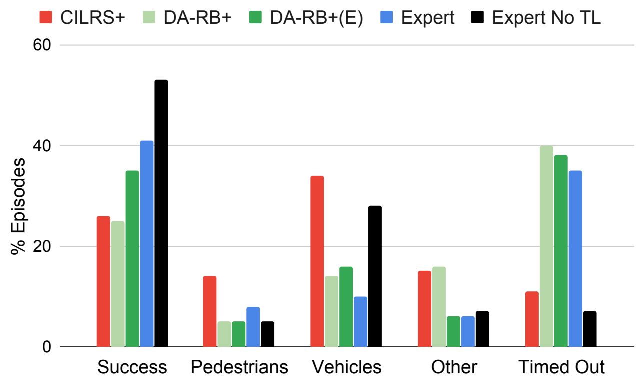


Figure 4: **Termination scenarios for different methods on dense setting of New Town & Weather condition.** We report the % episodes for success, collision with pedestrians, collision with vehicles, collision with other static objects and timed out scenarios for CILRS$^+$, DA-RB$^+$, DA-RB$^+$(E), Expert and Expert which ignores traffic lights (Expert No TL).

planner followed by a PID controller which outputs the controls. Since the expert policy has global information about all dynamic agents and access to a map of the town, the possibility of an error arising from endogenous factors in minimized. This ensures that the major source of variation in the behavior of the expert is due to the traffic lights and a minor source arising from the non-deterministic behavior of other dynamic agents. The results are reported in Fig. 4. We observe that 'Expert No TL' has a higher success rate compared to Expert which is counter-intuitive. This happens because the timed out scenarios decrease significantly resulting in an increase in success rate as well as collisions. This shows that following traffic lights is a huge catalyst leading to time outs and the evaluation is biased against the policies which drive cautiously. A better alternative to counter this issue can be to compute the time limit for each episode as a function of the performance of the expert policy, e.g., setting the time limit as 1.5 or 2 times the time taken by the expert policy to complete the episode. Furthermore, we also observe that performance of our approach (DA-RB$^+$(E)) is very similar to that of the Expert indicating that our approach enables the policy to learn appropriate driving behavior for urban driving.

5.2. Performance Metrics

The predominant metric used for reporting performance on the NoCrash Benchmark is success rate [4, 17, 19]. However, from Fig. 4 and Fig. 10 we observe that success rate is very deceptive and does not accurately represent the performance in dense urban scenarios. Moreover, it can be argued that safety and collision related metrics are more important than success rate in the presence of high density of dynamic agents in urban environments. We argue, given our findings, that success rate should always be used in conjunction with safety, collision and intervention related metrics while reporting driving performance. Also, the current evaluation setting in the NoCrash benchmark does not consider traffic light violation as a failure case. This is undesirable since traffic light violations can lead to fatal accidents and endanger the safety of other agents. Hence, this aspect needs to be appropriately represented. The Traffic-School benchmark [19] is a great first step in this direction.

5.3. Distribution of Weathers in the Evaluation Setting

CARLA 0.8.4 has 14 weather conditions, 4 of which are used for collecting training data (Weather ID: 1,3,6,8) and 2 of them are designated as new weather conditions (Weather ID: 10, 14) in the evaluation setting. However, we observe that these 2 new weather conditions do not reflect the overall trend exhibited by all 10 new weather conditions (Fig. 10, Fig. 11 and Fig. 5). Moreover, the metric scores are computed on a scale of 100 resulting in the performance difference on the 2 new weather conditions getting amplified. This may lead to misleading results when comparing performances on training and new weather conditions.

Potential alternatives include comparing the results on individual weather conditions, or assigning equal number of weathers to both training and the new weather setting. Furthermore, from Fig. 10, we observe that the 10 new weather conditions can be clearly divided into 2 categories - Weather ID: 2, 4, 5, 9, 11 and Weather ID: 7, 10, 12, 13, 14, based on the empirical performance of different methods with the latter being more difficult than the former. Hence, to accurately represent this distribution of new weathers, an appropriate evaluation setting should comprise of at least 2 weather conditions from each of

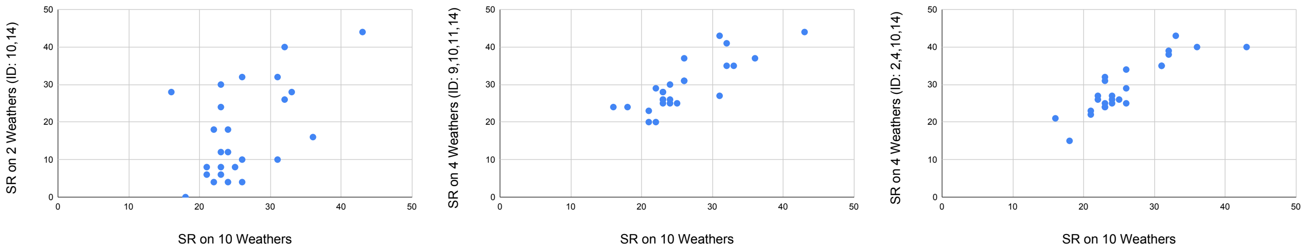


Figure 5: **Comparison of success rate (SR) on 2 new weathers against success rate on 4 new weathers on the dense setting of New Town & Weather.** We plot the correlation of (Left) SR on 2 new weathers (ID: 10, 14) with SR on 10 new weathers, (Middle) SR on 4 new weathers (ID: 9, 10, 11, 14) with SR on 10 new weathers, (Right) SR on 4 new weathers (ID: 2, 4, 10, 14) with SR on 10 new weathers, for all the methods shown in Figure 2 of the main paper.

the 2 categories, along with the 4 training weathers. To validate this claim, we compare the performance on 2 new weathers (ID: 10, 14) of the NoCrash benchmark against 4 new weathers with 2 weathers from each category described above and plot the correlation with the performance on all 10 new weathers (Fig. 5). From the plot, we observe that performance on 4 weathers correlates better with the performance on 10 weathers, thereby leading to a fair and more meaningful comparison.

5.4. Variance due to Training Seeds

High variance due to random training seeds is a widely acknowledged problem in the field of sensorimotor control [1, 4, 12, 18] which results in highly unstable policies. However, this aspect is not incorporated in the NoCrash benchmark with most of the approaches reporting the results using the best seed which may lead to misleading insights. Although, we conduct a comparative study on variance due to training seeds (Section 4.3 and Table 4 of the main paper) for CILRS$^+$, DAgger$^+$ and DA-RB$^+$, we observe (from Fig. 11) that some of values are ambiguous, e.g., CILRS$^+$ and DAgger$^+$ Iter 3 have similar standard deviation scores (Table 4 of the main paper) but from Fig. 11 we can clearly see that DAgger$^+$ Iter 3 has significantly higher standard deviation. This happens because the standard deviation values in Table 4 of the main paper are computed over mean of the success rate on 10 weather conditions which averages out the deviations. Therefore, while reporting variance and standard deviation, it is important to conduct a weather-wise analysis to provide an accurate estimate.

In Fig. 12, we also report the mean (M_{std}) and standard deviation (SD_{std}) computed over the standard deviations observed on all 10 individual weather conditions due to 5 random training seeds for each of the termination scenario. The goal of this analysis is to examine the robustness of the driving policy on each individual weather condition and identify a general trend across all 10 new weathers. A low value of M_{std} and SD_{std} indicates that the performance of the driving policy is stable across all weather conditions with high certainty, which is the most desirable scenario. Moreover, it is apparent from Fig. 12 that DA-RB$^+$ performs better than CILRS$^+$ and DAgger$^+$ and leads to stable driving policies.

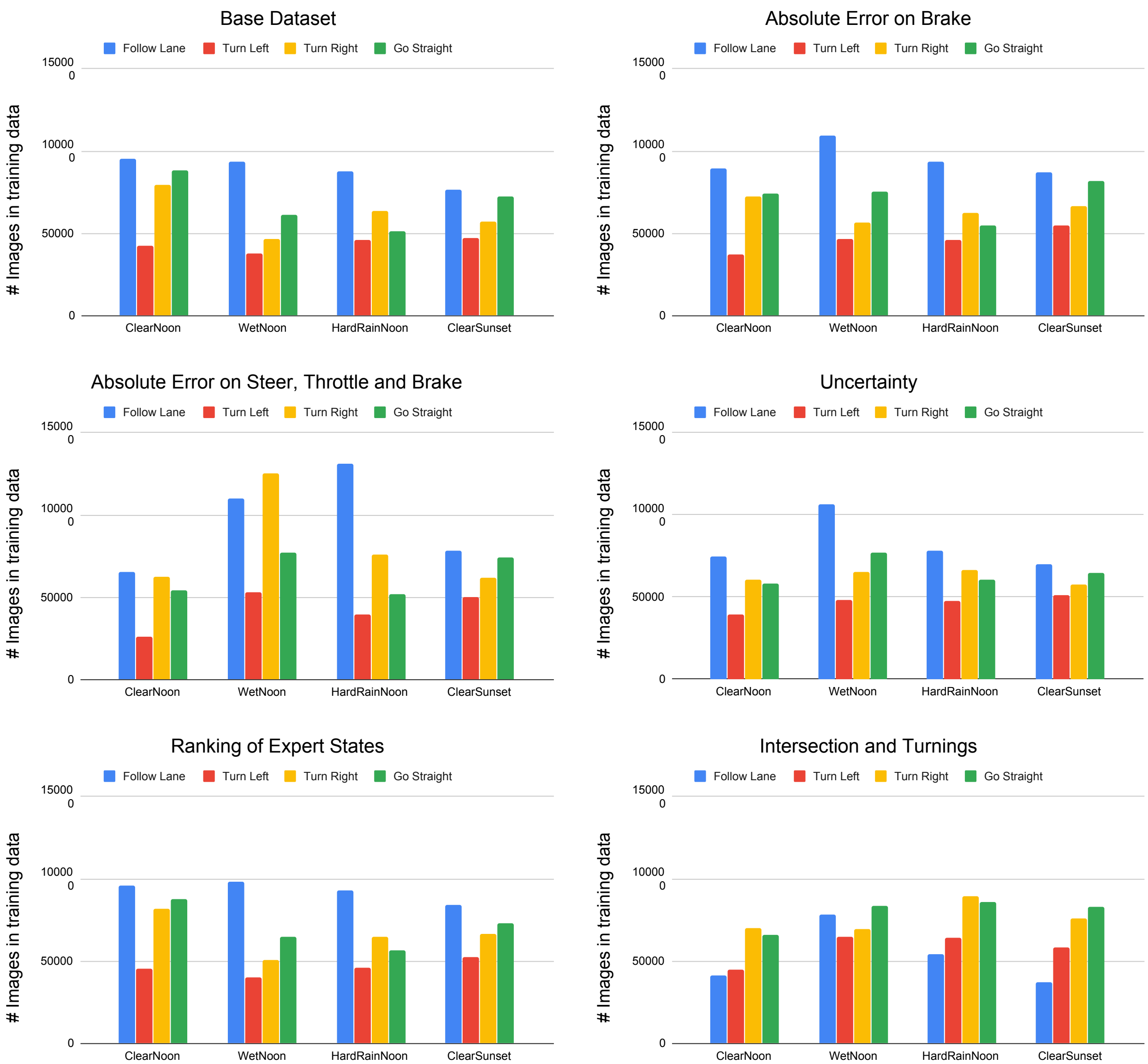


Figure 6: **Weather-wise data distribution over high level navigational commands induced by different sampling methods.** We report the number of images in the training data for the sampling methods described in Section 3.3, 4.4 and Table 5 of the main paper - base 10 hours dataset (Base), absolute error on brake (AE_b), absolute error on steer, throttle and brake (AE_{all}), uncertainty-based sampling (Unc), ranking of expert states (Rank) and intersection & turning scenarios (IT).

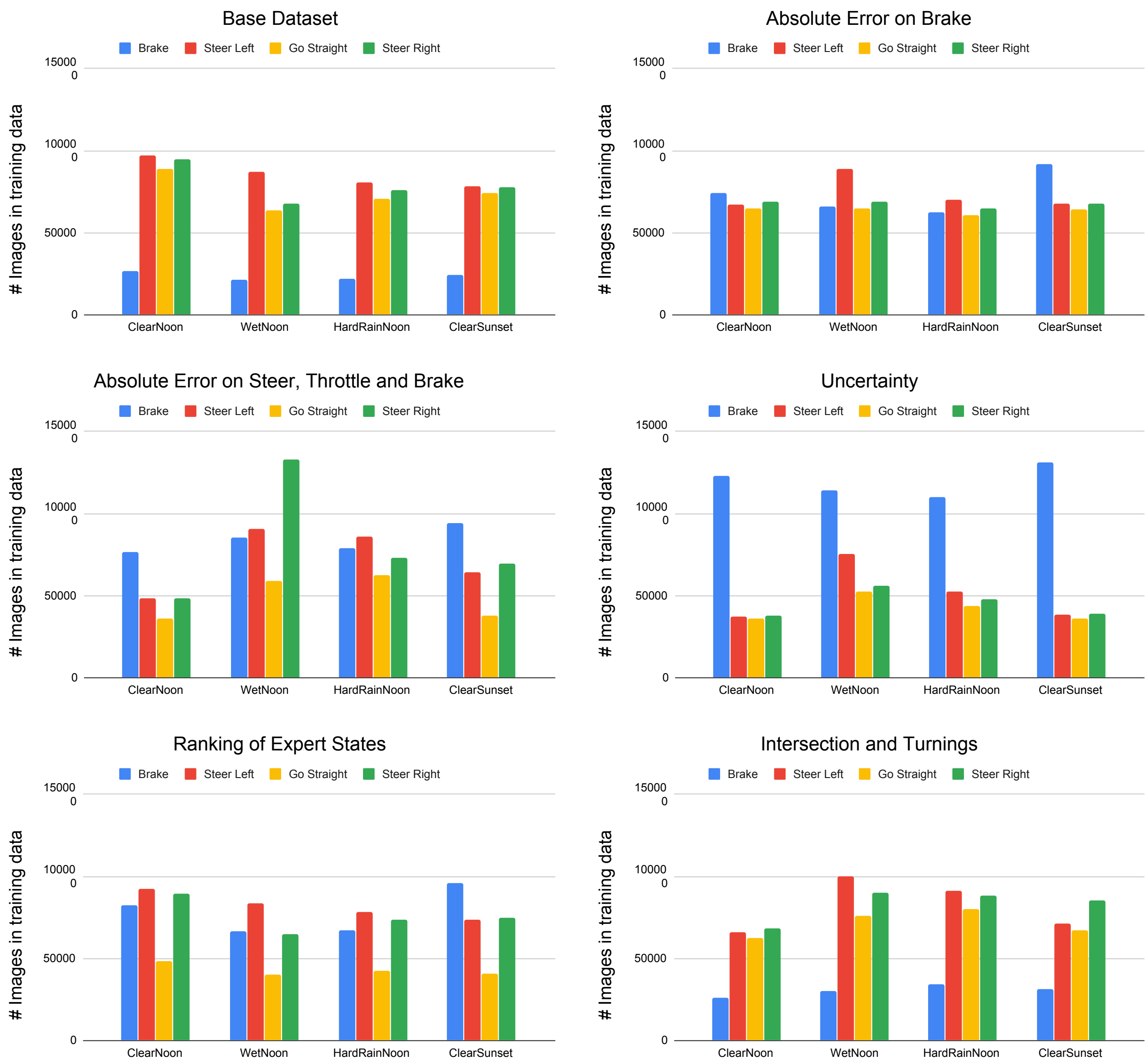


Figure 7: **Weather-wise data distribution over control categories induced by different sampling methods.** We report the number of images in the training data for the sampling methods described in Section 3.3, 4.4 and Table 5 of the main paper - base 10 hours dataset (Base), absolute error on brake (AE_b), absolute error on steer, throttle and brake (AE_{all}), uncertainty-based sampling (Unc), ranking of expert states (Rank) and intersection & turning scenarios (IT).

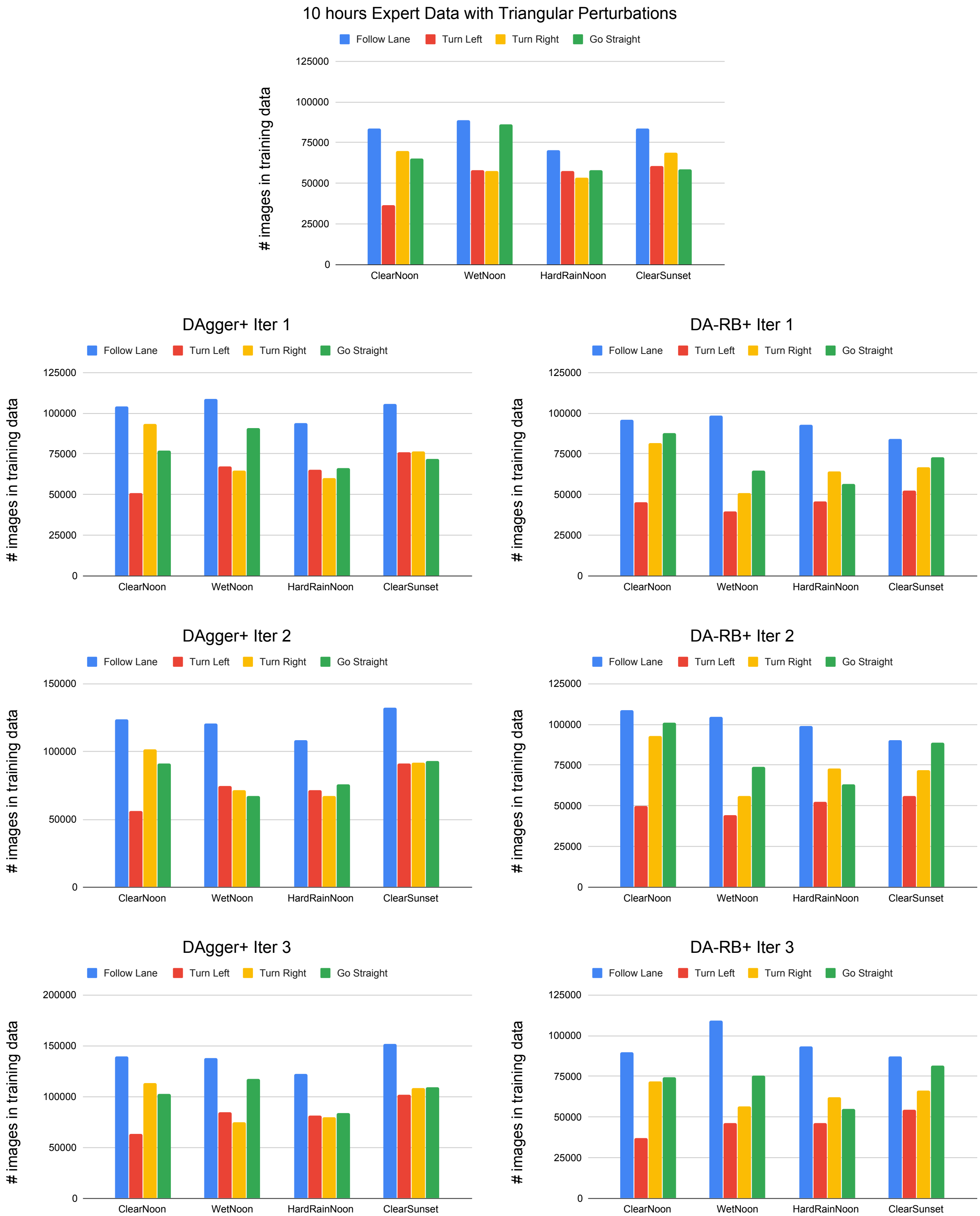


Figure 8: **Weather-wise data distribution over high level navigational commands induced by DAgger$^+$ and DA-RB$^+$.** We report the number of images in the training data for CILRS$^+$ (10 hours expert data with triangular perturbations) and multiple iterations of DAgger$^+$ and DA-RB$^+$.

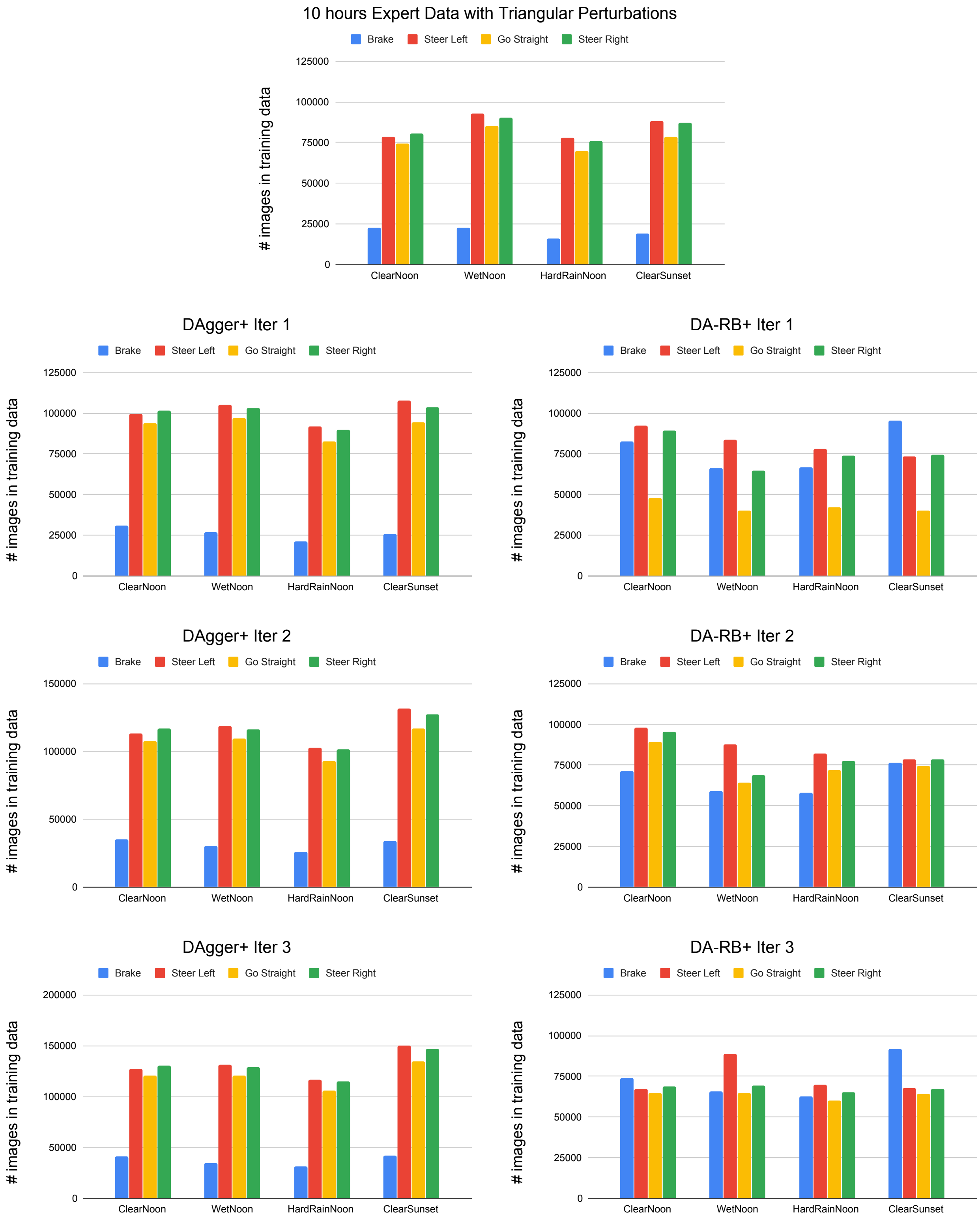


Figure 9: **Weather-wise data distribution over control categories induced by DAgger^{+} and DA-RB^{+}.** We report the number of images in the training data for CILRS^{+} (10 hours expert data with triangular perturbations) and multiple iterations of DAgger^{+} and DA-RB^{+}.

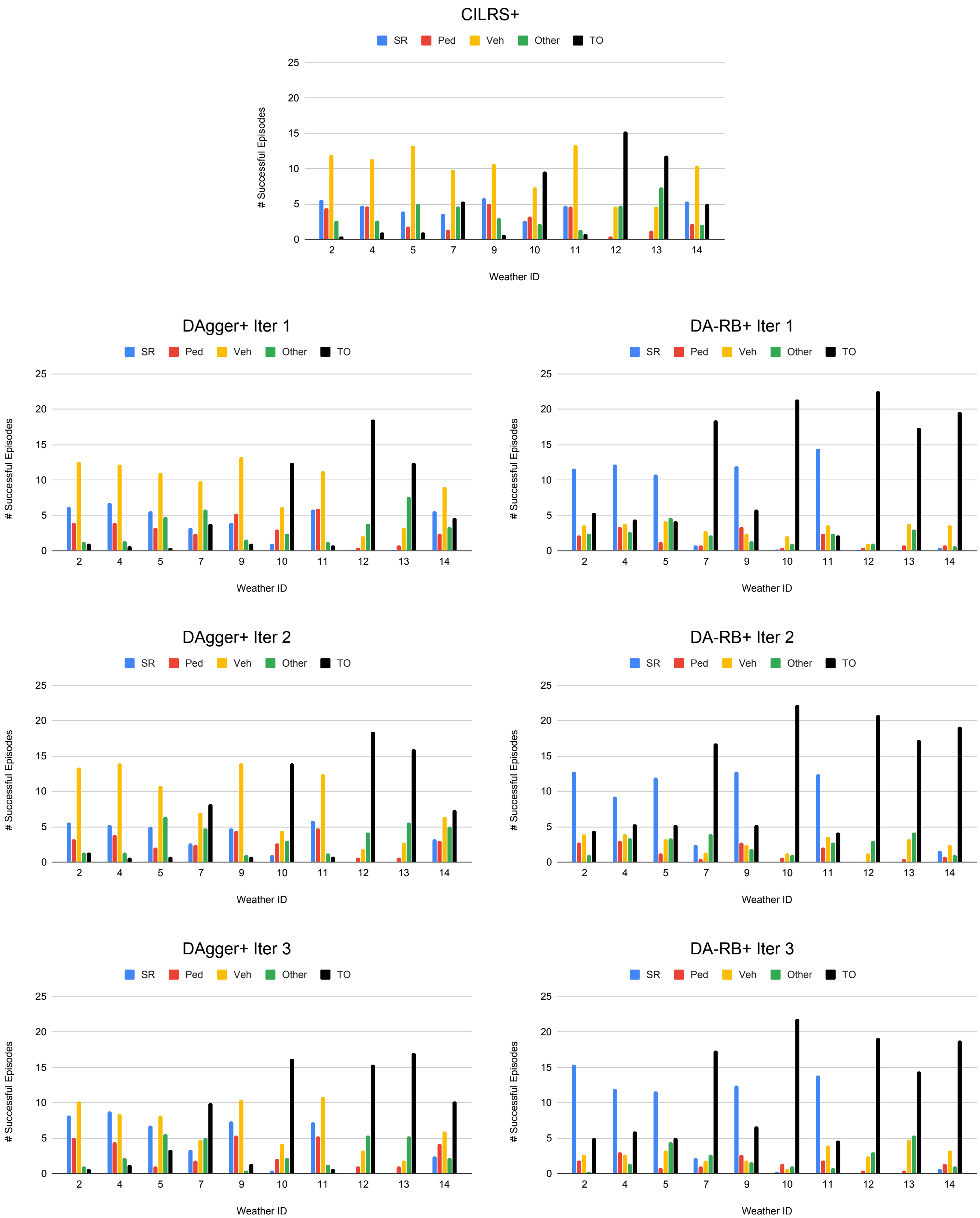


Figure 10: **Weather-wise mean of different termination scenarios wrt. 5 random training seeds on the dense setting of New Town & Weather.** We report the weather-wise mean (out of 25 episodes) of success rate (SR), collision with pedestrians (Ped), vehicles (Veh), other static objects (Other) & timed out (TO) scenarios for CILRS^{+}, multiple iterations of DAgger^{+} and DA-RB^{+}. Weather ID - 2:CloudyNoon, 4:WetCloudyNoon, 5:MidRainNoon, 7:SoftRainNoon, 9:CloudySunset, 10:WetSunset, 11:WetCloudySunset, 12:MidRainSunset, 13:HardRainSunset, 14:SoftRainSunset

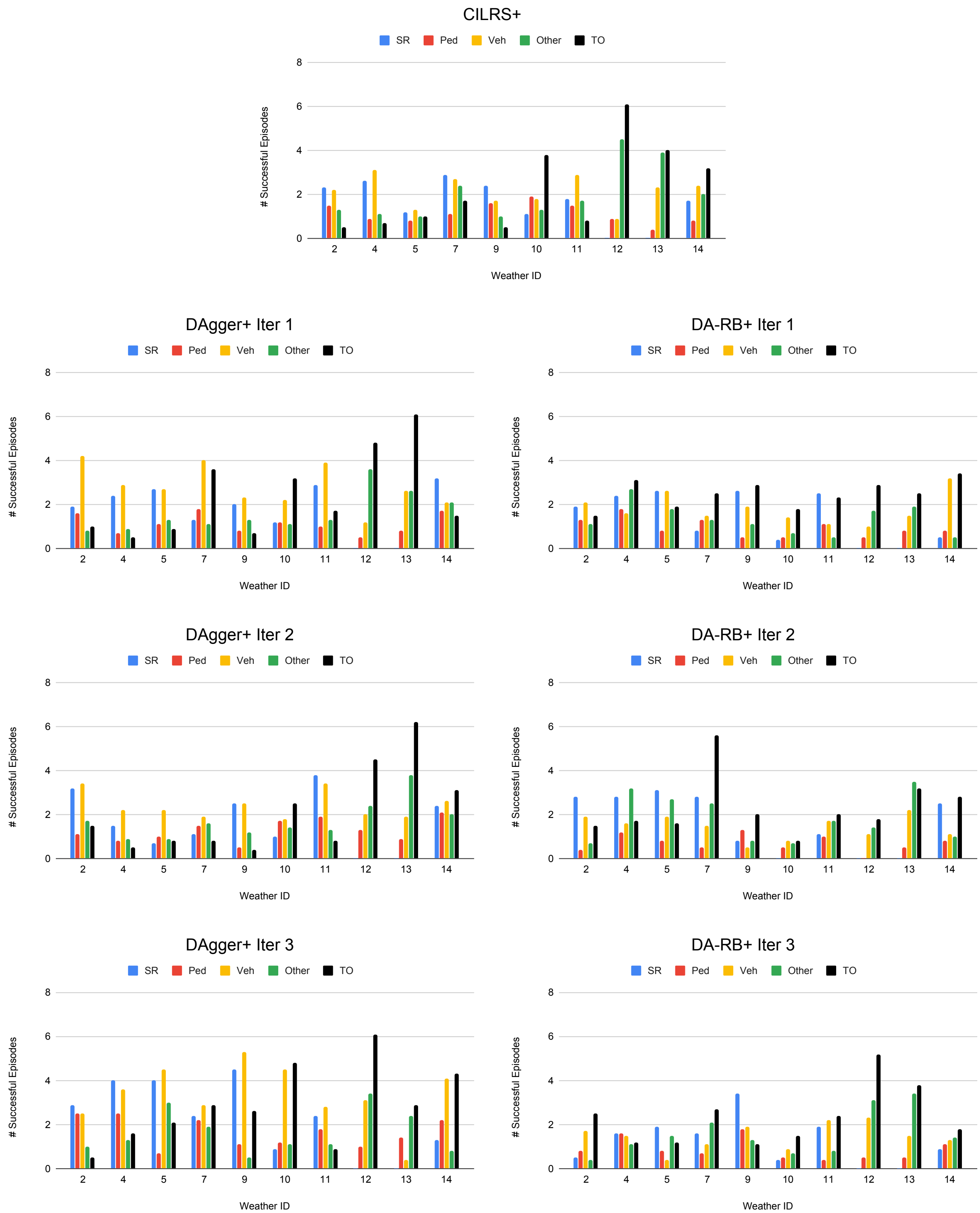
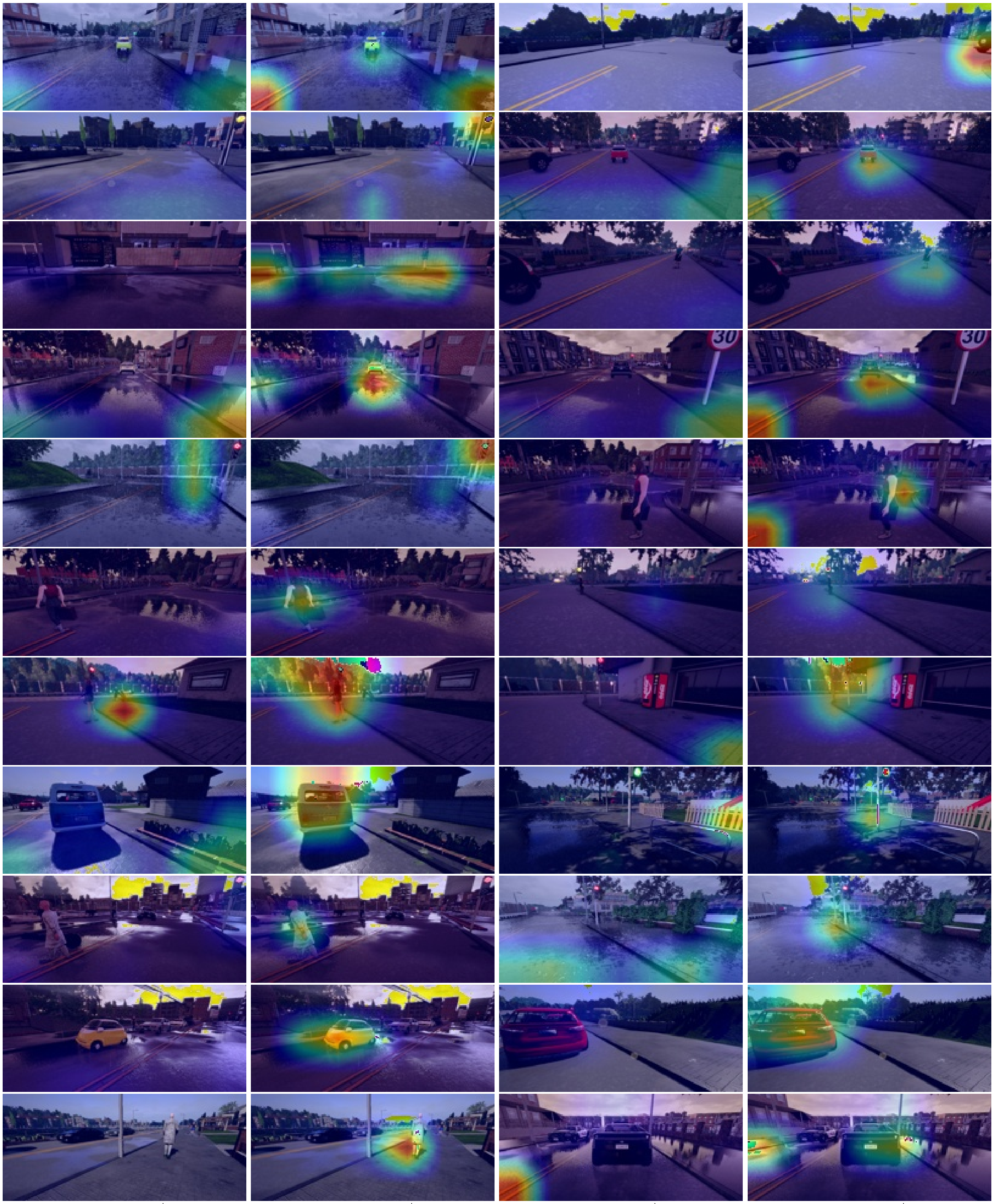


Figure 11: **Weather-wise standard deviation of different termination scenarios wrt. 5 random training seeds on the dense setting of New Town & Weather.** We report the weather-wise standard deviation (out of 25 episodes) of success rate (SR), collision with pedestrians (Ped), vehicles (Veh), other static objects (Other) & timed out (TO) scenarios for CILRS$^+$, multiple iterations of DAgger$^+$ and DA-RB$^+$. Weather ID - 2:CloudyNoon, 4:WetCloudyNoon, 5:MidRainNoon, 7:SoftRainNoon, 9:CloudySunset, 10:WetSunset, 11:WetCloudySunset, 12:MidRainSunset, 13:HardRainSunset, 14:SoftRainSunset

Figure 12: **Mean and standard deviation computed on the standard deviations of each termination scenario wrt. 5 random training seeds on the dense setting of New Town & Weather.** Left: Mean over the standard deviations of all 10 new weathers, Right: Standard deviation over the standard deviations of all 10 new weathers.



(a) CILRS$^+$ (b) DA-RB$^+$ (c) CILRS$^+$ (d) DA-RB$^+$

Figure 13: **GradCAM Attention Maps.**

References

[1] Marcin Andrychowicz, Dwight Crow, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 9

[2] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 3

[3] Zalán Borsos, Sebastian Curi, Kfir Yehuda Levy, and Andreas Krause. Online variance reduction with mixtures. In *Proc. of the International Conf. on Machine learning (ICML)*, 2019. 3

[4] Felipe Codevilla, Eder Santana, Antonio M. López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 1, 2, 3, 4, 7, 8, 9

[5] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proc. Conf. on Robot Learning (CoRL)*, 2017. 1

[6] Elmar Haussmann, Michele Fenzi, Kashyap Chitta, Jan Ivanecky, Hanson Xu, Donna Roy, Akshita Mittel, Nicolas Koumchatzky, Clement Farabet, and Jose M. Alvarez. Scalable Active Learning for Object Detection. In *Proc. IEEE Intelligent Vehicles Symposium (IV)*, 2020. 6

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[8] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 4

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2015. 3

[10] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *arXiv.org*, 1906.08158, 2019. 6

[11] Michael Laskey, Jonathan Lee, Roy Fox, Anca D. Dragan, and Ken Goldberg. DART: noise injection for robust imitation learning. In *Proc. Conf. on Robot Learning (CoRL)*, 2017. 6

[12] Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv.org*, 1810.08591, 2018. 9

[13] Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010. 3, 5

[14] Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011. 3, 5

[15] Nataniel Ruiz, Samuel Schulter, and Manmohan Chandraker. Learning to simulate. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019. 4

[16] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2017. 7

[17] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M. López. Multimodal end-to-end autonomous driving. *arXiv.org*, 1906.03199, 2019. 8

[18] Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end simulated driving. In *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2017. 9

[19] Albert Zhao, Tong He, Yitao Liang, Haibin Huang, Guy Van den Broeck, and Stefano Soatto. Lates: Latent space distillation for teacher-student driving policy learning. *arXiv.org*, 1912.02973, 2019. 8