A. Multimodal Unsupervised Image-to-image Translation (MUNIT) Framework

In this section, we give a brief description of the Multimodal Unsupervised Image-to-Image Translation (MUNIT) framework proposed by Huang et al. (2018).

MUNIT is based on a partially shared latent space assumption. It first assumes that the latent space of images can be decomposed into a content space and a style space. Then, it further assumes that images in different domains share a common content space but not the style space. In our context, the domains are different environmental types (e.g., sunny and rainy weather types) and the images are driving scenes. Therefore, the content space captures information that is shared by scenes of different environmental types, such as the shapes of roads and the roadside trees. The style space captures the variants of visual representation of the same content from different environmental types, for instance, the unique degrees of illumination, amounts of rain, and cloud patterns for the rainy weather type. Intuitively, the corresponding style space of an environmental type can be used to represent the environmental condition space of the environmental type.

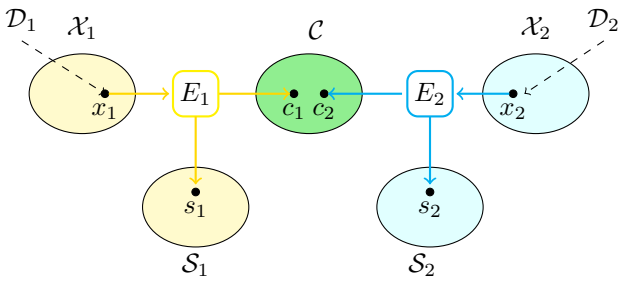

Figure 7. Structure of MUNIT. Images in each domain $\mathcal{X}_i$ are encoded to a shared content space $\mathcal{C}$ and a domain-specific style space $\mathcal{S}_i$ through an encoder E_i. Each encoder has an inverse generator G_i omitted from this figure. For each domain $\mathcal{X}_i$, there is also a discriminator D_i which detect whether the image belong to $\mathcal{X}_i$.

Figure 7 presents the structure of MUNIT, where $\mathcal{X}_1$ and $\mathcal{X}_2$ denote two different domains (i.e., two different environmental types in our context). For each domain $\mathcal{X}_i$, there are an encoder E_i which projects the images from $\mathcal{X}_i$ to a domain-invariant content space $\mathcal{C}$ and a domain-specific style space $\mathcal{S}_i$, and a decoder G_i which reproduces images of $\mathcal{X}_i$. Furthermore, D_1 and D_2 are two discriminators that detect whether the image belongs to $\mathcal{X}_1$ or $\mathcal{X}_2$, respectively. Specifically, these discriminators are excepted to differentiate whether the input images are real or synthetic ones (i.e., images produced by a well-trained generator). All E_i, G_i, and D_i in a MUNIT model are incarnated as deep neural networks, and learned by optimising the loss function that is consisting of the following costs:

- **Image Reconstruction Loss:** minimising the loss of image reconstruction for each $< E_i, G_i >$, which encourages reconstruction in the direction: image→latent→image.
- **Latent Reconstruction Loss:** minimising the loss of latent code reconstruction for each $< E_i, G_i >$, which encourages reconstruction in the direction: latent → image → latent.
- **GAN Loss:** achieving the equilibrium point in the minimax game for each $< G_i, D_i >$, where D_i aims at distinguishing between synthetic images produced by G_i and real images in $\mathcal{X}_i$.

In general, the total loss is defined as:

$$\min_{E_1,E_2,G_1,G_2} \max_{D_1,D_2} \mathcal{L}_{\mathsf{GAN}_1} + \mathcal{L}_{\mathsf{GAN}_2} + \lambda_x(\mathcal{L}^{\mathsf{image}}_{\mathsf{recon}_1} + \mathcal{L}^{\mathsf{image}}_{\mathsf{recon}_2}) + \lambda_c(\mathcal{L}^{\mathsf{content}}_{\mathsf{recon}_1} + \mathcal{L}^{\mathsf{content}}_{\mathsf{recon}_2}) + \lambda_s(\mathcal{L}^{\mathsf{style}}_{\mathsf{recon}_1} + \mathcal{L}^{\mathsf{style}}_{\mathsf{recon}_2})$$

where $\mathcal{L}^{\mathsf{image}}_{\mathsf{recon}_i}$ represents the image reconstruction loss for $< E_i, G_i >$. $\mathcal{L}^{\mathsf{content}}_{\mathsf{recon}_i}$ and $\mathcal{L}^{\mathsf{style}}_{\mathsf{recon}_i}$ respectively represents the content code reconstruction loss and the style code reconstruction loss for $< E_i, G_i >$. $\mathcal{L}_{\mathsf{GAN}_i}$ denotes the GAN loss for $< G_i, D_i >$. λ_x, λ_c, λ_s are weights that control the importance of reconstruction terms.

Based on a well-trained MUNIT model, we can easily translate an image to another domain. For example, to translate an image $x_1 \in \mathcal{X}_1$ to $\mathcal{X}_2$, the MUNIT model first decomposes it into a content latent code c_1 and a style latent code s_1 by

the encoder E_1, i.e., $(c_1, s_1) = E_1(x_1)$. Then it uses the generator G_2 to produce an image x_2 by combining the content code c_1 and a style latent code s_2 sampled from the style space $\mathcal{S}_2$ of $\mathcal{X}_2$, i.e., $x_2 = G_2(c_1, s_2)$. By sampling different style latent code s_2 from the style space $\mathcal{S}_2$ of $\mathcal{X}_2$, multiple images with different appearances can be produced. For example, as shown in the left half of Figure 4, we can use one original driving scene to synthesise driving scenes under different environmental conditions of an environmental condition type. The original driving scene belongs to the sunny weather type. When transforming the original scene to the environmental type of 'Rain', we sampled two different style codes from the corresponding style space of the 'Rain' type. We can see that both synthesised driving scenes belong to the same environmental type but they have observable different degrees of light and amounts of rain. Besides this transformation, we can also use MUNIT to transform the entire set of driving scenes under different environmental conditions into those under the same environmental condition by using the same style code, as shown in the right half of Figure 4. We applied the same style code from the style space of the environmental type of 'Rain' on two different original driving scenes. It can be observed that the synthesised scenes are not only under the same environmental type, but also the same environmental condition by having similar degrees of light and amounts of rain.

B. Details for Datasets and Target DNN-based ADSs

In this section, we present details about the subject DNN-based ADSs and datasets used in our experiments.

B.1. Subject DNN-based ADSs

We focus on testing DNN-based ADSs that perform end-to-end steering angle control. Three popular pre-trained DNN-based ADSs that have been widely used in previous work (Pei et al., 2017; Tian et al., 2018; Zhang et al., 2018; Zhou et al., 2020), i.e., Dave-orig (0bserver07, 2016), Dave-dropout (Navoshta, 2016), and Chauffeur (Emef, 2016), are selected as the subject systems. Dave-orig and Dave-dropout both use the CNN models base on the DAVE-2 self-driving architecture from NVIDIA (Bojarski et al., 2016). Chauffeur is one of the top-ranked DNN models in the Udacity self-driving car challenge (Udacity, 2016). It consists of one CNN model and one LSTM model. For an input driving scene, the CNN first extracts features from the input and then the LSTM predicts the steering angle of the input based on the concatenation of 100 features extracted by the CNN from the previous 100 consecutive driving scenes.

Table 4. Details of datasets

DATASETS	NUMBER	DURATION	ENV. TYPE
UDACITY TRAINING	33808	-	SUNSHINE
UDACITY TESTING	5614	-	SUNSHINE
LOS ANGELES - NIGHT DRIVE (UTAH, 2019)	2000	1:17:03	NIGHT
DRIVING ON CALIFORNIA FREEWAY (TOURS, 2018)	1600	38:49	SUNSHINE
RAIN ON A CAR ROOF (VIDS, 2014)	1000	1:09:04	RAIN
DRIVING ON SNOW - GREENVILLE (MCGOWAN, 2018)	1200	28:56	SNOW IN DAYTIME
DRIVING IN THE SNOW (BADVBOYNOFEAR, 2018)	1200	1:04:53	SNOW IN NIGHT

B.2. Datasets

Table 4 presents the detailed information of all datasets used in our experiments. All three DNNs are trained on the Udacity dataset (Udacity, 2016), which is for the Udacity self-driving car challenge, containing 33808 training samples and 5614 testing samples. Each sample consists of a driving scene captured by a camera mounted behind the windshield of a driving car and the simultaneous steering angle issued by the human driver.

We study five environmental types (night, sunshine, rain, snow in daytime and snow in night) that are representatives of the runtime environments for DNN-based ADSs. The acquisition of the datasets used for training the MUNIT models is simple: we just searched videos longer than 20 minutes of driving in the five environmental types from YouTube, and conducted an automatic downsampling on the selected video to skip consecutive frames having similar contents and used the retained frames to construct the datasets. Table 4 shows the names of the videos and their links as in citations, the number of sampled frames, the duration of the videos, and the corresponding environmental types.

C. Further Experimental Details

In this section, we present further experimental details as well as some in-depth analysis.

C.1. Comparison with Baselines on Effectiveness

In order to evaluate the effectiveness of TACTIC, we compare TACTIC against two baseline methods, i.e., an approach using randomly sampled environmental conditions $\mathrm{R_c}$ and the state-of-the-art DeepRoad.

For TACTIC, we respectively execute TACTIC with two coverage-guiding strategies: KMNC (denoted as TACTIC$^{\mathrm{KMNC}}$) and NBC (denoted as TACTIC$^{\mathrm{NBC}}$) on each of the three subject DNN-based ADSs, under the five environmental types. In total, 30 experimental cases are executed (2 strategies $\times$ 5 environmental types $\times$ 3 systems). Furthermore, to reduce the randomness of (1+1) ES used in TACTIC, we conduct 10 runs for each case and average the results.

For $\mathrm{R_c}$, for each subject DNN-based ADS, we randomly sample 4 style vectors (equal to the number of critical style vectors generated by TACTIC) for each of the environmental types. In total, 15 experimental cases are executed (5 environmental types $\times$ 3 systems). We also repeat $\mathrm{R_c}$ 10 times for each experimental case and average the results.

For DeepRoad, we use the training datasets of the MUNIT models in TACTIC for training DeepRoad, and then generate driving scenes for each of the environmental types. Moreover, given the Udacity testing dataset, DeepRoad can only generate the same number (5614) of testing driving scenes as the Udactiy dataset for each environmental type, since it transforms each driving scene into the other in a deterministic way. Therefore, for TACTIC, we apply just one critical style on the Udactiy testing dataset each time and then average the results for a fair comparison.

Table 5. Average results of KMNC-guided TACTIC using (1+1) ES

MODEL	ENV. TYPE	COVERAGE		NUMBER OF ERRORS			
		KMNC	NBC	10°	20°	30°	40°
DAVE-ORIG	NIGHT	69.12%	32.15%	21277.6	19499.2	15505.3	8952.7
	SUNSHINE	52.81%	9.90%	5668.1	540.5	22.8	0.4
	RAIN	46.21%	4.88%	13482.0	7084.2	3058.4	809.3
	SNOW IN DAYTIME	47.08%	3.65%	17938.7	6596.3	737.5	109.9
	SNOW IN NIGHT	72.95%	33.80%	21415.2	19480.5	17721.2	13677.6
DAVE-DROPOUT	NIGHT	51.55%	21.10%	16547.8	4846.6	917.5	113.4
	SUNSHINE	38.54%	10.32%	7031.0	793.5	30.0	0.2
	RAIN	40.07%	11.03%	13334.3	4707.7	454.6	67.8
	SNOW IN DAYTIME	39.82%	11.79%	18828.0	17709.0	11744.8	1966.2
	SNOW IN NIGHT	33.06%	8.48%	21539.9	18721.5	4701.6	7.3
CHAUFFEUR	NIGHT	78.23%	38.03%	1608.8	117.9	0.0	0.0
	SUNSHINE	69.41%	17.03%	439.1	16.0	0.0	0.0
	RAIN	61.56%	8.02%	15342.6	8377.9	1507.3	0.8
	SNOW IN DAYTIME	72.93%	20.23%	19188.2	5275.1	145.9	0.5
	SNOW IN NIGHT	71.87%	24.62%	9995.8	2040.4	25.8	0.0

Comparison results with $\mathrm{R_c}$. Table 5 and Table 6 summarise the average results achieved by TACTIC$^{\mathrm{KMNC}}$ and TACTIC$^{\mathrm{NBC}}$ over 10 repeated runs for each experimental case, respectively. Table 7 summarises the average results achieved by TACTIC$^{\mathrm{KMNC}}$ and TACTIC$^{\mathrm{NBC}}$ over 10 repeated runs for each experimental case. In Table 5, Table 6, and Table 7, Column *Coverage* is the achieved test coverage and Column *Number of Errors* is the number of detected erroneous behaviours. From these results. we have the following findings:

In terms of the achieved coverage, TACTIC achieves higher coverage than $\mathrm{R_c}$ in most cases. For Dave-orig, Dave-dropout, and Chauffeur, TACTIC$^{\mathrm{KMNC}}$ achieves on average 12.33%, 4.03%, and 5.87% higher coverage than $\mathrm{R_c}$, respectively, and TACTIC$^{\mathrm{NBC}}$ achieves on average 15.69%, 6.38%, and 7.42% higher coverage than $\mathrm{R_c}$, respectively. In terms of the detected erroneous behaviours, TACTIC detects more erroneous behaviours than $\mathrm{R_c}$. In total, for Dave-orig, Dave-dropout, and Chauffeur TACTIC$^{\mathrm{KMNC}}$ detects on average 470.77%, 1495.11%, and 738.80% more erroneous behaviours than $\mathrm{R_c}$, and TACTIC$^{\mathrm{NBC}}$ detects on average 330.47%, 1171.31%, and 501.72% more erroneous behaviours than $\mathrm{R_c}$. Furthermore, we also observe that, when the error bound increases, $\mathrm{R_c}$ hardly detects erroneous behaviours while TACTIC still shows a strong

Table 6. Average results of NBC-guided TACTIC using (1+1) ES

MODEL	ENV. TYPE	COVERAGE		NUMBER OF ERRORS			
		KMNC	NBC	10°	20°	30°	40°
DAVE-ORIG	NIGHT	73.68%	35.92%	18675.1	13790	8627.7	4303.7
	SUNSHINE	56.24%	13.81%	2629.8	196.5	22.7	0.7
	RAIN	50.71%	7.30%	9795.2	3479.7	1993.8	687.7
	SNOW IN DAYTIME	54.41%	8.67%	13450.8	3396.8	845.1	209.5
	SNOW IN NIGHT	72.09%	33.30%	20879.0	17978.1	14561.1	12074.7
DAVE-DROPOUT	NIGHT	52.46%	22.09%	11114.4	2967.8	463.2	22.7
	SUNSHINE	43.26%	16.12%	1879.2	251.3	15.3	0.1
	RAIN	43.88%	13.75%	12872.3	5872.6	583.7	79.5
	SNOW IN DAYTIME	42.77%	13.92%	17648.9	14779.5	8012.4	930.0
	SNOW IN NIGHT	33.47%	8.55%	20973.3	8882.6	1831.0	21.1
CHAUFFEUR	NIGHT	78.93%	39.92%	2965.9	198.7	0.0	0.0
	SUNSHINE	69.85%	20.49%	306.3	22.8	0.0	0.0
	RAIN	62.19%	11.21%	12560.4	6627.6	915.5	3.4
	SNOW IN DAYTIME	71.36%	23.90%	12600.5	669.4	73.0	0.2
	SNOW IN NIGHT	72.61%	26.97%	2955.5	240.2	12.8	0.0

ability to detect erroneous behaviours. For example, for Dave-dropout, $\mathrm{R_c}$ detects no erroneous behaviours when the error bound is set to $40°$ in the environmental type of "Snow in Daytime", while TACTIC$^{\mathrm{KMNC}}$ and TACTIC$^{\mathrm{NBC}}$ still detect on average 1966.2 and 930 erroneous behaviours, respectively.

Following existing guidelines (Arcuri & Briand, 2014), to further investigate whether the differences between TACTIC and $\mathrm{R_c}$ are statistically significant, we use the nonparametric pairwise Mann-Whitney U test (Capon, 1991) and Vargha-Delaney Statistics $\hat{A}_{12}$ (Vargha & Delaney, 2000), where for two approaches A and B, A has significantly better performance than B if $\hat{A}_{12}$ is higher than 0.5 and the p-value is less than 0.05. The level of significance (α) is set to 0.05.

Table 8 reports the statistical test results comparing TACTIC$^{\mathrm{KMNC}}$ and $\mathrm{R_c}$, while those for comparing TACTIC$^{\mathrm{KMNC}}$ and $\mathrm{R_c}$ are reported in Table 9. The results show that: (1) for the achieved test coverage, TACTIC obtains significantly higher coverage than $\mathrm{R_c}$ in 86.67% (26 out of 30) experimental cases ($\hat{A}_{12}$ greater than 0.78 and p-value less than 0.01). For the other 4 cases, TACTIC also achieves comparable coverage to $\mathrm{R_c}$, with an average of only 0.66% lower coverage than $\mathrm{R_c}$. (2) for the number of detected erroneous behaviours, TACTIC detects significantly more erroneous behaviours than $\mathrm{R_c}$ for 91.67% (110 cases) of the 120 cases (4 error bounds and each bound has 30 experimental cases), by having $\hat{A}_{12}$ greater than 0.75 and p-value less than 0.04. For the 10 exceptional cases where the significance is not observed, the numbers of erroneous behaviours detected by both TACTIC and $\mathrm{R_c}$ are not large enough to pass the significance tests, but in no case, TACTIC detects fewer erroneous behaviours than $\mathrm{R_c}$.

In summary, we can conclude that TACTIC manages to detect more diverse and more serious erroneous behaviours than $\mathrm{R_c}$.

Comparison results with DeepRoad. Table 10 presents the results achieved by DeepRoad. Table 11 and Table 12 present the average results achieved by TACTIC$^{\mathrm{KMNC}}$ and TACTIC$^{\mathrm{NBC}}$, respectively. Recall that we separately apply each critical environmental condition produced by TACTIC and average the results to conduct a comparison with DeepRoad.

In terms of the detected erroneous behaviours, the comparison results are consistent with the results of comparing TACTIC with $\mathrm{R_c}$ that TACTIC always detects substantially more erroneous behaviours than DeepRoad on all three subject systems with the four error bounds. In total, for Dave-orig, Dave-dropout, and Chauffeur, TACTIC$^{\mathrm{KMNC}}$ detects on average 390.29%, 956.14%, and 514.88% more erroneous behaviours than DeepRoad, respectively, and TACTIC$^{\mathrm{NBC}}$ detects on average 341.92%, 931.90%, and 416.74% more erroneous behaviours than DeepRoad, respectively.

The comparison of the achieved coverage is more complex. For Dave-orig and Dave-dropout, TACTIC shows slightly higher coverage than DeepRoad, which, specifically, is that TACTIC$^{\mathrm{KMNC}}$ achieves on average 1.52% and 0.82% higher coverage, and TACTIC$^{\mathrm{NBC}}$ achieves on average 4.53% and 1.59% higher coverage than DeepRoad, respectively. For Chauffeur, the average coverage achieved by TACTIC is lower than DeepRoad, however, the difference is very small, which is an average of 0.71% between TACTIC$^{\mathrm{KMNC}}$ and DeepRoad, and 1.05% between TACTIC$^{\mathrm{NBC}}$ and DeepRoad.

Table 7. Average results of Using Random Styles ($\mathbf{R_c}$)

Model	Env. Type	Coverage		Number of Errors			
		KMNC	NBC	10°	20°	30°	40°
Dave-orig	Night	43.55%	3.18%	2971.2	269.7	25.5	1.9
	Sunshine	43.34%	1.97%	1300.5	103.8	4.7	0.0
	Rain	42.04%	2.00%	1484.5	44.9	0.1	0.0
	Snow in Daytime	47.15%	3.88%	3605.2	197.7	10.6	0.8
	Snow in Night	52.04%	10.10%	7323.4	2583.0	749.7	72.5
Dave-dropout	Night	36.70%	8.90%	965.2	42.1	3.1	0.0
	Sunshine	39.62%	12.51%	635.8	57.0	3.4	0.0
	Rain	33.07%	8.18%	878.6	66.4	4.0	0.0
	Snow in Daytime	37.83%	10.15%	701.3	57.3	5.5	0.0
	Snow in Night	31.68%	7.85%	2260.6	61.7	5.0	0.0
Chauffeur	Night	73.01%	20.21%	345.1	18.3	0.0	0.0
	Sunshine	68.28%	13.99%	294.5	11.2	0.0	0.0
	Rain	62.16%	10.18%	1048.2	163.2	0.0	0.0
	Snow in Daytime	63.05%	8.10%	3126.7	178.9	7.1	0.1
	Snow in Night	67.44%	16.77%	665.8	6.4	0.0	0.0

The reason is that the environmental condition to transform each testing driving scene generated by DeepRoad is independent of each other, whereas TACTIC uses one critical style vector corresponding to the same environmental condition to transform all the driving scenes. Consequently, the driving scenes generated by DeepRoad are under independent environmental conditions, and therefore, may result in higher coverage in some cases. This is especially the case for Chauffeur, since its CNN is only used as a feature extractor (cf. Appendix B), rendering the test coverage of Chauffeur more sensitive to diverse environmental conditions. Nevertheless, it is more important to study the critical environmental conditions instead of the various random ones, since most environmental conditions can be well handled by these popular DNN-based ADSs, as confirmed by the experimental results.

C.2. Ablation Study

To justify the selection of (1+1) ES in TACTIC, we implement a new version of TACTIC by replacing (1+1) ES with the random search (RS) and compare the effectiveness of the two versions. Except for the search algorithm, the other implementation and settings, e.g., the fitness function, are not changed. We denote TACTIC using RS with KMNC-guided as RS^{KMNC} and the one with NBC-guided as RS^{NBC} in the experiments. For each experimental case, we also conduct 10 runs and average the results to reduce the randomness of the random search.

Table 13 and Table 14 summarise the average results achieved by RS^{KMNC} and RS^{NBC} over 10 repeated runs for each experimental case, respectively. We discuss the comparison results from two aspects, i.e., the achieved test coverage and the number of detected erroneous behaviours. In terms of the achieved coverage, for Dave-orig, Dave-dropout, and Chauffeur, $\text{TACTIC}^{\text{KMNC}}$ achieves on average 7.34%, 3.88%, and 4.35% higher coverage than RS^{KMNC}, respectively, and $\text{TACTIC}^{\text{NBC}}$ achieves on average 9.98%, 4.21%, and 4.14% higher coverage than RS^{NBC}, respectively. In terms of the detected erroneous behaviours, in total, $\text{TACTIC}^{\text{KMNC}}$ detects on average 105.73%, 428.12%, and 120.24% more erroneous behaviours than RS^{KMNC}, and $\text{TACTIC}^{\text{NBC}}$ detects on average 62.98%, 619.51%, and 155.80% more erroneous behaviours than RS^{NBC} for Dave-orig, Dave-dropout, and Chauffeur, respectively. Additionally, we also find that the difference between the number of erroneous behaviours detected by the two approaches increases with the error bound. For example, when setting the error bound to 40°, for Dave-dropout, RS^{KMNC} and RS^{NBC} only respectively detect on average 0.1 and 0.2 erroneous behaviours in the environmental type of "Snow in Daytime", while $\text{TACTIC}^{\text{KMNC}}$ and $\text{TACTIC}^{\text{NBC}}$ respectively detect on average 1966.2 and 930 erroneous behaviours.

Furthermore, Table 15 and Table 16 respectively report the statistical test results comparing TACTIC using (1+1) ES and TACTIC using RS with two coverage-guiding strategies. From these tables, we can observe that: (1) In terms of the achieved coverage, $\text{TACTIC}^{\text{KMNC}}$ achieves significantly higher coverage than RS^{KMNC} for 73.33% (11 cases) of the 15 cases, and

TACTIC$^{\text{NBC}}$ achieves significantly higher coverage than RS$^{\text{NBC}}$ for 80% (12 cases) of the 15 cases, by having $\hat{A}_{12}$ greater than 0.7 and p-value less than 0.04. (2) In terms of the detected erroneous behaviours, TACTIC$^{\text{KMNC}}$ detects significantly more erroneous behaviours than RS$^{\text{KMNC}}$ for 91.67% (55 cases) of the 60 cases (4 error bounds and each bound has 15 experimental cases), and TACTIC$^{\text{NBC}}$ detects significantly more erroneous behaviours than RS$^{\text{NBC}}$ for 93.33% (56 cases) of the 60 cases, by having $\hat{A}_{12}$ greater than 0.76 and p-value less than 0.02. For the few exceptional cases, both TACTIC using (1+1) ES and using RS detect few erroneous behaviours.

Therefore, we can conclude that the critical environmental conditions produced by TACTIC using (1+1)ES can achieve higher test coverage and detect more serious erroneous behaviours than those produced by TACTIC using RS.

Table 8. Statistical test results for TACTIC^{KMNC} and R_c.

MODEL	METRIC	NIGHT		SUNSHINE		RAIN		SNOW IN DAYTIME		SNOW IN NIGHT	
		P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$
DAVE-ORIG	KMNC	0.00009	1	0.00009	1	0.00085	0.92	0.36667	0.45	0.00009	1
	NBC	0.00009	1	0.00008	1	0.00008	1	0.22433	0.395	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00009	1	0.00008	1	0.00004	1	0.00008	1	0.00009	1
	40°	0.00009	1	0.03879	0.65	0.00003	1	0.00005	1	0.00008	1
DAVE-DROPOUT	KMNC	0.00009	1	0.07023	0.7	0.00009	1	0.00021	0.97	0.00009	1
	NBC	0.00009	1	0.10606	0.33	0.00008	1	0.00008	1	0.01835	0.78
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00008	1	0.00007	1	0.00005	1	0.00008	1	0.00008	1
	40°	0.00003	1	0.18406	0.55	0.00003	1	0.00003	1	0.00746	0.75
CHAUFFEUR	KMNC	0.00009	1	0.00363	0.86	0.00008	0	0.00009	1	0.00009	1
	NBC	0.00009	1	0.00009	1	0.00009	0	0.00009	1	0.00009	1
	10°	0.00009	1	0.00043	0.945	0.00009	1	0.00009	1	0.00009	1
	20°	0.00008	1	0.00432	0.85	0.00009	1	0.00009	1	0.00009	1
	30°	FAILED		FAILED		0.00003	1	0.00006	1	0.00003	1
	40°	FAILED		FAILED		0.03893	0.65	0.31317	0.545	FAILED	

* We highlight the cases where TACTIC^{KMNC} is worse than R_c with red.
** FAILED represents that the number of erroneous behaviours detected by the two approaches are not large enough to pass the significance tests.

Table 9. Statistical test results for TACTIC^{NBC} and R_c

MODEL	METRIC	NIGHT		SUNSHINE		RAIN		SNOW IN DAYTIME		SNOW IN NIGHT	
		P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$
DAVE-ORIG	KMNC	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	NBC	0.00009	1	0.00008	1	0.00008	1	0.00009	1	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00009	1	0.00008	1	0.00004	1	0.00008	1	0.00009	1
	40°	0.00006	1	0.01742	0.7	0.00003	1	0.00005	1	0.00009	1
DAVE-DROPOUT	KMNC	0.00009	1	0.00009	1	0.00009	1	0.00016	0.98	0.00009	1
	NBC	0.00009	1	0.00008	1	0.00008	1	0.00009	1	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00054	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00008	1	0.0088	0.815	0.00005	1	0.00008	1	0.00008	1
	40°	0.00003	1	0.18406	0.55	0.00003	1	0.00003	1	0.00037	0.9
CHAUFFEUR	KMNC	0.00009	1	0.00021	0.97	0.41023	0.535	0.00009	1	0.00009	1
	NBC	0.00009	1	0.00008	1	0.10614	0.67	0.00009	1	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00008	1	0.00054	1	0.00009	1	0.00009	1	0.00009	1
	30°	FAILED		FAILED		0.00003	1	0.00006	1	0.00298	0.8
	40°	FAILED		FAILED		0.00298	0.8	0.33505	0.54	FAILED	

* We highlight the cases where TACTIC^{NBC} is worse than R_c with red.
** FAILED represents that the number of erroneous behaviours detected by the two approaches are not large enough to pass the significance tests.

Table 10. Results of DeepRoad

Model	Env. Type	Coverage		Number of Errors			
		KMNC	NBC	10°	20°	30°	40°
Dave-orig	Night	40.99%	5.99%	613	114	33	9
	Sunshine	40.97%	2.05%	355	40	3	0
	Rain	40.23%	2.05%	487	20	0	0
	Snow in Daytime	45.21%	3.81%	1250	90	2	0
	Snow in Night	55.66%	21.50%	3189	1996	802	173
Dave-dropout	Night	33.06%	9.28%	402	29	1	0
	Sunshine	34.48%	10.31%	201	16	1	0
	Rain	31.97%	8.46%	241	21	1	0
	Snow in Daytime	34.99%	10.24%	182	13	1	0
	Snow in Night	31.29%	8.81%	2324	62	1	0
Chauffeur	Night	69.19%	18.64%	140	6	0	0
	Sunshine	65.08%	11.16%	86	5	0	0
	Rain	62.25%	11.30%	232	35	0	0
	Snow in Daytime	63.64%	9.55%	876	188	23	0
	Snow in Night	68.17%	20.43%	573	57	0	0

Table 11. Results of TACTIC^{KMNC} comparing with DeepRoad

Model	Env. Type	Coverage		Number of Errors			
		KMNC	NBC	10°	20°	30°	40°
Dave-orig	Night	48.78%	13.41%	5398.5	5013.3	4199.0	2704.5
	Sunshine	42.27%	4.23%	1812.5	182.0	7.3	0.0
	Rain	41.75%	3.17%	2631.5	1330.8	389.3	85.5
	Snow in Daytime	40.45%	2.29%	4404.0	1518.8	231.5	30.5
	Snow in Night	55.05%	22.22%	5342.5	4976.8	4492.5	3456.5
Dave-dropout	Night	36.58%	12.14%	4418.5	1268.3	211.3	22.3
	Sunshine	34.35%	9.00%	1813.8	216.8	6.3	0.0
	Rain	35.24%	9.51%	1751.0	393.5	53.0	6.5
	Snow in Daytime	34.46%	9.51%	4221.3	3506.5	1803.8	243.5
	Snow in Night	31.72%	8.63%	5421.8	4690.5	1106.0	1.8
Chauffeur	Night	69.05%	19.47%	380.3	22.8	0.0	0.0
	Sunshine	64.94%	10.94%	109.0	2.8	0.0	0.0
	Rain	61.28%	7.53%	3887.5	2199.0	353.3	0.0
	Snow in Daytime	65.37%	10.86%	4865.3	1340.5	36.5	0.0
	Snow in Night	66.65%	16.27%	2550.0	569.3	8.5	0.0

Table 12. Results of TACTIC^{NBC} comparing with DeepRoad

MODEL	ENV. TYPE	COVERAGE		NUMBER OF ERRORS			
		KMNC	NBC	10°	20°	30°	40°
DAVE-ORIG	NIGHT	54.59%	21.72%	4885.0	3898.3	2662.5	1620.5
	SUNSHINE	45.37%	6.62%	708.8	41.5	4.8	0.0
	RAIN	40.71%	2.90%	3035.8	1422.5	816.5	304.5
	SNOW IN DAYTIME	41.80%	3.34%	3708.0	820.5	154.3	47.3
	SNOW IN NIGHT	60.03%	26.64%	5041.0	4418.0	3716.3	2843.0
DAVE-DROPOUT	NIGHT	39.97%	14.89%	2116.8	774.8	124.5	5.3
	SUNSHINE	33.93%	11.05%	281.0	14.0	0.8	0.0
	RAIN	34.71%	9.61%	3796.5	1921.8	186.9	28.0
	SNOW IN DAYTIME	35.57%	9.76%	4872.3	4122.5	2688.5	342.3
	SNOW IN NIGHT	30.54%	8.77%	5306.0	2433.5	580.0	11.3
CHAUFFEUR	NIGHT	68.25%	19.95%	370.3	41.3	0.0	0.0
	SUNSHINE	64.80%	11.17%	73.3	5.3	0.0	0.0
	RAIN	61.34%	7.65%	4146.8	2198.3	369.5	2.5
	SNOW IN DAYTIME	63.76%	10.69%	3042.3	160.5	16.8	0.0
	SNOW IN NIGHT	66.37%	14.97%	569.4	59.5	3.2	0.0

Table 13. Average eesults of KMNC-guided TACTIC using Random Search

MODEL	ENV. TYPE	COVERAGE		NUMBER OF ERRORS			
		KMNC	NBC	10°	20°	30°	40°
DAVE-ORIG	NIGHT	52.48%	10.24%	15471.7	4710.5	411.7	0.0
	SUNSHINE	43.66%	2.26%	2364.9	142.6	9.4	0.1
	RAIN	42.55%	2.06%	3629.9	225.9	5.8	0.2
	SNOW IN DAYTIME	50.48%	5.43%	10496.0	1232.4	48.1	0.8
	SNOW IN NIGHT	64.56%	25.42%	19629.1	15729.0	10319.7	4325.2
DAVE-DROPOUT	NIGHT	38.10%	9.58%	4907.7	328.6	36.4	0.0
	SUNSHINE	38.03%	10.92%	1429.5	47.7	1.3	0.0
	RAIN	34.83%	8.58%	1599.1	135.6	11.0	0.2
	SNOW IN DAYTIME	37.36%	9.99%	2180.0	209.6	12.0	0.1
	SNOW IN NIGHT	32.04%	8.50%	18887.9	4175.9	13.9	0.0
CHAUFFEUR	NIGHT	73.75%	22.29%	1202.5	54.7	0.0	0.0
	SUNSHINE	68.83%	16.51%	407.7	9.8	0.0	0.0
	RAIN	61.76%	8.64%	3616.4	309.5	0.0	0.0
	SNOW IN DAYTIME	65.45%	9.99%	12976.1	600.0	1.6	0.0
	SNOW IN NIGHT	70.25%	20.99%	3475.6	253.1	0.0	0.0

Table 14. Average results of NBC-guided TACTIC using Random Search

MODEL	ENV. TYPE	COVERAGE		NUMBER OF ERRORS			
		KMNC	NBC	10°	20°	30°	40°
DAVE-ORIG	NIGHT	53.09%	10.73%	15833.3	4783.0	384.0	9.1
	SUNSHINE	45.48%	2.85%	1451.4	81.2	5.0	0.0
	RAIN	42.35%	2.00%	3633.9	188.1	3.8	0.0
	SNOW IN DAYTIME	51.87%	6.80%	9730.0	748.8	20.0	14.3
	SNOW IN NIGHT	65.45%	25.69%	19341.5	14917.3	9955.4	4966.7
DAVE-DROPOUT	NIGHT	38.61%	10.51%	4497.2	340.6	32.6	0.0
	SUNSHINE	42.56%	15.56%	902.4	42.8	1.1	0.0
	RAIN	35.08%	8.94%	1056.7	86.2	1.7	0.0
	SNOW IN DAYTIME	42.17%	13.86%	975.5	111.0	9.9	0.2
	SNOW IN NIGHT	32.38%	8.51%	18267.9	4321.6	8.1	0.0
CHAUFFEUR	NIGHT	75.00%	27.53%	691.8	32.1	0.0	0.0
	SUNSHINE	69.39%	18.39%	296.3	15.2	0.0	0.0
	RAIN	62.29%	11.82%	2397.7	232.7	0.0	0.0
	SNOW IN DAYTIME	66.60%	10.99%	12336.4	337.8	0.0	0.0
	SNOW IN NIGHT	70.40%	23.59%	2461.3	137.2	0.0	0.0

Table 15. Statistical test results for TACTIC^{KMNC} and RS^{KMNC}.

MODEL	METRIC	NIGHT		SUNSHINE		RAIN		SNOW IN DAYTIME		SNOW IN NIGHT	
		P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$
DAVE-ORIG	KMNC	0.00009	1	0.00009	1	0.00455	0.85	0.00009	0	0.00009	1
	NBC	0.00009	1	0.00009	1	0.00007	1	0.00009	0	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	40°	0.00009	1	0.13903	0.605	0.00004	1	0.00005	1	0.00009	1
DAVE-DROPOUT	KMNC	0.00009	1	0.07923	0.7	0.00009	1	0.00454	0.85	0.00009	1
	NBC	0.00009	1	0.10606	0.33	0.00009	1	0.00009	1	0.42505	0.53
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00008	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00008	1	0.00007	1	0.00009	1	0.00009	1	0.00008	1
	40°	0.00009	1	0.18406	0.55	0.00007	1	0.00009	1	0.00746	0.75
CHAUFFEUR	KMNC	0.00009	1	0.04808	0.725	0.00921	0.185	0.00009	1	0.0001	0.99
	NBC	0.00009	1	0.00256	0.875	0.00009	0	0.00009	1	0.00009	1
	10°	0.00029	0.96	0.00628	0.835	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00008	0.92	0.00009	1	0.00009	1	0.00009	1
	30°	FAILED		FAILED		0.00009	1	0.00009	1	0.00009	1
	40°	FAILED		FAILED		0.03893	0.65	0.08403	0.6	FAILED	

* We highlight the cases where TACTIC^{KMNC} is worse than RS^{KMNC} with red.

** FAILED represents that the number of erroneous behaviours detected by the two approaches are not large enough to pass the significance tests.

Table 16. Statistical test results for TACTIC^{NBC} and RS^{NBC}.

MODEL	METRIC	NIGHT		SUNSHINE		RAIN		SNOW IN DAYTIME		SNOW IN NIGHT	
		P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$	P-VALUE	$\hat{A}_{12}$
DAVE-ORIG	KMNC	0.00009	1	0.00009	1	0.00009	1	0.00021	0.97	0.00009	1
	NBC	0.00009	1	0.00008	1	0.00009	1	0.00009	1	0.00009	1
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	30°	0.00009	1	0.00008	1	0.00008	1	0.00009	1	0.00009	1
	40°	0.00009	1	0.01742	0.7	0.00003	1	0.00009	1	0.00009	1
DAVE-DROPOUT	KMNC	0.00009	1	0.14486	0.645	0.00009	1	0.23633	0.6	0.00009	1
	NBC	0.00009	1	0.00009	1	0.00009	1	0.33873	0.44	0.43989	0.525
	10°	0.00009	1	0.00009	1	0.00009	1	0.00009	1	0.00009	1
	20°	0.00009	1	0.00012	0.99	0.00009	1	0.00009	1	0.00009	1
	30°	0.00008	1	0.00014	0.98	0.00007	1	0.00009	1	0.00008	1
	40°	0.00003	1	0.18406	0.55	0.00004	1	0.00009	1	0.00037	0.9
CHAUFFEUR	KMNC	0.00009	1	0.00862	0.82	0.27252	0.415	0.00009	1	0.00009	1
	NBC	0.00009	1	0.00009	1	0.42504	0.53	0.00009	1	0.00009	1
	10°	0.00009	1	0.02459	0.765	0.00009	1	0.00506	0.845	0.03201	0.75
	20°	0.00009	1	0.00018	0.975	0.00009	1	0.00009	1	0.00009	1
	30°	FAILED		FAILED		0.00003	1	0.00003	1	0.00298	0.8
	40°	FAILED		FAILED		0.00298	0.8	0.08374	0.6	FAILED	

* We highlight the cases where TACTIC^{NBC} is worse than RS^{NBC} with red.
** FAILED represents that the number of erroneous behaviours detected by the two approaches are not large enough to pass the significance tests.