Scaling-Aware Data Selection for End-to-End Autonomous Driving Systems

Supplementary Material

7. Experiment Protocols

7.1. Dataset and Virtual Clip Creation

We conduct experiments using the `Navtrain` [14] and `trainval` splits of `OpenScene` [11] as the combined training and pool datasets. `OpenScene` is a redistribution of the NuPlan dataset [19], subsampled to 2 Hz, and contains approximately 120 hours of driving data with dense annotations. The `Navtrain` split is curated within the NAVSIM framework [14] by filtering out trivial driving scenarios from the `trainval` split of `OpenScene`. In both experiments, evaluation is performed on `navtest` [14], a validation set curated analogously from the `test` split of `OpenScene`.

The `Navtrain` and `trainval` splits consist of 1,192 and 1,250 individual driving sessions, respectively, with durations ranging from 30 seconds to 50 minutes. In addition to this large temporal variation, the total number of available driving sessions remains limited, and treating individual frames as independent samples to would be both unrealistic and inconsistent with the temporal structure of driving data. Hence, to have more samples to work with and to align our data handling with common industry practice [16] we segment each driving log into fixed-length *virtual clips* of 10 seconds (corresponding to 20 frames at 2 Hz). Below, we describe how we create virtual clips are created.

Table 4. Train-pool clip counts for OpenScene and Navtrain

	Train	Pool
Openscene	1000	31539
Navtrain	460	4141

We segment each driving log into fixed-length *virtual clips* of 10 seconds. Given the dataset's sampling rate of 2 Hz, each virtual clip contains 20 frames. For each log, non-overlapping clips are extracted sequentially from the start of the log, and any remaining portion shorter than 10 seconds is discarded. For example, a 23-second log yields two clips covering [0–10) s and [10–20) s, while the final 3 seconds are omitted. Following this procedure, the `Navtrain` split yields a total of 4,601 virtual clips after discarding 11,268 out of 103,288 frames (10.9%). For `OpenScene`, we obtain 32,539 virtual clips, with 12,086 out of 662,866 frames (1.8%) omitted due to incomplete segments. The train-pool clip counts are summarized in Table 4

7.2. Training details

As already mentioned in the main body of the paper, we use the Hydra-MDP model [33] that won the NAVSIM benchmark in 2024 [14] by a significant margin. In addition to the imitation trajectory loss, the model distills the rule-compliance scores of each trajectory, obtained with prior simulations. We initialize the model's encoder as pretrained VoVNetV2-99 backbone [31, 40]. In accordance with the training recipe of provided in the paper [33], we use Adam optimizer [29] without any weight decay and keep the learning rate fixed throughout the training. We set the per-GPU batch size to 20.

In the `Navtrain` experiments, all runs are conducted using 8×A100 GPUs with a learning rate of $1e{-4}$. Each experiment is repeated with three random seeds (0, 2025, 424242), and the reported results are averaged over these runs, with the standard deviation shown as a subscript. The base experiment and budgets up to 800 clips are trained for 60 epochs, while the 1,600- and 2,400-clip settings are trained for 50 and 45 epochs, respectively, to reduce compute cost.

For the `OpenScene` experiments, the rule-compliance distillation losses are disabled due to their high computational overhead needed to run intensive simulations to calculate those scores. All runs use 16×A100 GPUs with a learning rate of $2e{-4}$ and a fixed training length of 40 epochs. Experiments with 2,000, 4,000, and 8,000 clips are repeated with two seeds (0, 2025), while smaller-budget runs use three seeds (0, 2025, 424242) to ensure stability.

7.3. Details of the Baselines

Random. For each budget B, Random selection is constructed from a single randomized ordering of the pool. Specifically, we shuffle all clips once using a fixed seed (seed = 42) and define the selected set for budget B as the first B clips in this ordering. This ensures that selections for larger budgets are strict supersets of those for smaller budgets.

Uncertainty [24]. We score each pool clip by the entropy of its model-predicted trajectory logits. Let z_i denote the (pre-softmax) logits for sample x_i, and let $p_i = \mathrm{softmax}(z_i)$ be the corresponding probability distribution over candidate trajectories. The uncertainty score is taken as the Shannon entropy $H_i = -\sum_k p_{i,k} \log p_{i,k}$. The uncertainty score is calculated for each frame in the clip, and we simply take average of the frame uncertainty scores to aggregate it at the clip level. Clips with higher entropy correspond to more

ambiguous or uncertain model predictions and are therefore preferred. To construct a budget-B selection, we compute H_i for every pool item once, rank all items by entropy in descending order, and pick the top B. We share the procedure in Algorithm 2.

Algorithm 2 Entropy-Based Uncertainty Selection

Require: Pool samples $\{x_i\}$, model $f(\cdot)$, budget B
Ensure: Selected set S of size B
Initialize $S = \emptyset$
for each sample x_i in the pool **do**
  $z_i = f(x_i)$ ▷ trajectory logits
  $p_i = \mathrm{softmax}(z_i)$
  $H_i = -\sum_k p_{i,k} \log p_{i,k}$ ▷ entropy score
end for
Rank all pool samples by H_i in descending order
$S \leftarrow$ top-B samples under this ranking
return S

Coreset [46]. We adopt the standard geometric Coreset selection procedure shown in Algorithm 3. Starting from an initial set of training indices s^0, the algorithm iteratively adds the pool element that is farthest under the chosen distance measure $\Delta(\cdot, \cdot)$ from the current selected set. Specifically, we use Euclidean Distance. At each iteration, Coreset identifies the sample $u \in s^{pool}$ that maximizes the minimum distance to the existing set s, and then augments s with u. This expansion continues until the total size reaches $B + |s^0|$, yielding the Coreset of size B from the pool.

Algorithm 3 Coreset

Require: train sample indices s^0, budget B, pool indices s^{pool}
Initialize $s = s^0$
repeat
  $u = \arg\max_{i \in s^{pool}} \min_{j \in s} \Delta(x_i, x_j)$
  $s = s \cup \{u\}$
until $|s| = B + |s^0|$
return s

Chameleon [54]. Chameleon is a domain-mixture framework that relies on embeddings computed from the training domains. First, each cluster is embedded using representations from the base model's feature space. For each cluster, sample embeddings are averaged which produces one embedding per cluster. A cluster–cluster affinity matrix is then constructed using a kernel function applied to pairs of domain embeddings. Given this affinity matrix, Chameleon applies kernel ridge regression (KRLS) to compute a score S_i for each domain, reflecting how informative or influential that domain is relative to all others. Finally, the mixture weight for domain i is obtained by normalizing these scores with a softmax, $\alpha_i = \mathrm{softmax}(S_i)$, and data are sampled from domains according to these mixture weights. We use the pretraining mode in our experiments, as we re-train the model from scratch for each budget and we set the ridge parameter as $\lambda = 1$. The pseudo-code is provided in Algorithm 4.

Algorithm 4 Chameleon Domain Weighting (Pretraining Mode)

Require: Training clusters $\mathcal{D} = \{D_1, \ldots, D_k\}$, ridge parameter λ, embedding layer L, budget B
Ensure: Selected set S of size B
Extract domain embeddings:
  $x_i = \frac{1}{|D_i|} \sum_{a \in D_i} h_\theta^{(L)}(a)$ for each domain D_i
Construct feature matrix $X = [x_1^\top, \ldots, x_k^\top]$
Compute affinity matrix $\Omega_D = XX^\top$
Compute KRLS scores $S_\lambda(D_i)$ for each domain D_i using Ω_D
Compute domain weights:
  $\alpha_i^{PT} = \frac{\exp(S_\lambda^{-1}(D_i))}{\sum_{j=1}^k \exp(S_\lambda^{-1}(D_j))}$
Sample B points from domains according to mixture weights $\{\alpha_i^{PT}\}$
return S

8. More Results on the Experiments and Ablations

Due to the space constraints in the main body of the paper, we present more results here.

Experiments on `Openscene`. The full validation EPDMS and BRMR results for the Openscene experiments can be found in Table 6. The breakdown of the validation EPDMS subscores are shared in Table 9. The scaling curves obtained from different cities are shared in Figure 7.

Experiments on `Navtrain`. The full validation EPDMS and BRMR results for the Navtrain experiments can be found in Table 7. The breakdown of the validation EPDMS subscores are shared in Table 10. We also provide the city distributions induced by different method at various budgets in Figure 9. The scaling curves obtained from different cities are shared in Figure 8. The scaling curves obtained from different cities are shared in Figure 8.

Ablation with Caption-based Clustering. To generate the clip captions, we used the Qwen-2.5VL-32B-Instruct model with the following caption: "*This is a 10 second*

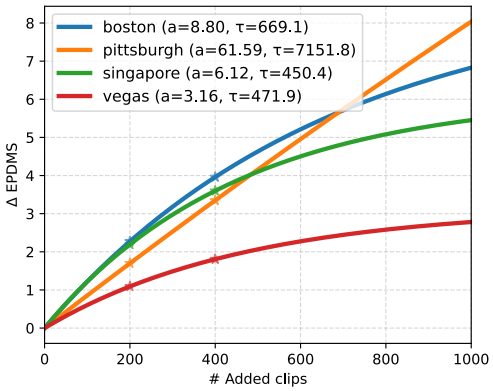


Figure 7. Performance scalings of different cities for the OpenScene experiment.

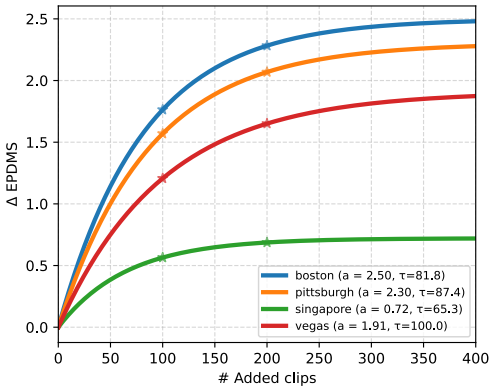
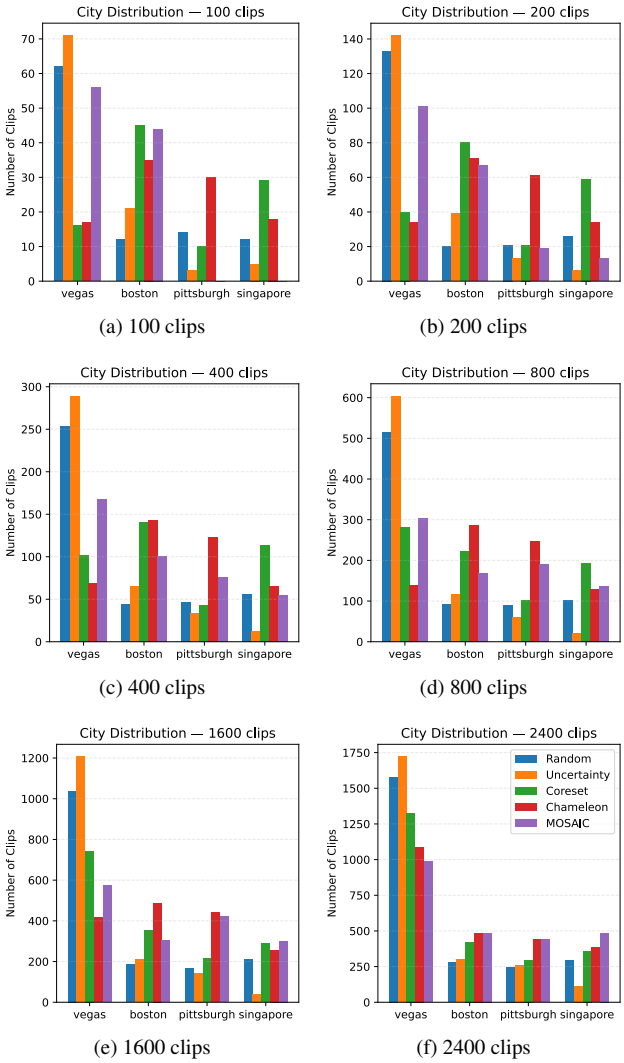


Figure 8. Performance scalings of different cities for the main Navtrain experiment.

long video of your student driving. The clip might include discontinuities, sudden changes in the driving environment. Describe the driving environment that your student is driving through and your student's driving actions. Please describe the driving condition including the location, weather, road users, and their motions. During your description, there are several things to keep in mind. 1. Please pay attention only to the objects on the driving roads and ignore the background. 2. Ignore the brands of the vehicles. 3. Describe it if objects are partially occluded by others, or are in areas with different brightness such as under shades. Please provide a concise description in one paragraph with less than 150 words. Do not mention anything that you are certain does not exist! No statements about uncertain objects or events (no 'maybe' or 'might' or 'possibly'). All responses must be in English only!"

On the generated clip captions, we extract TF–IDF features using the top 1,024 unigrams and bigrams after removing common English stop words. We then perform clustering in this TF–IDF space, forming six clusters. The dominant scene characteristics of each cluster are determined by their highest-weight unigrams and bigrams, as summarized in Table 3. We additionally conduct a qualitative assessment of the resulting groups and confirm that the clusters are coherent and semantically meaningful.

In fact, we have first attempted using "sentence-transformers/all-mpnet-base-v2" model downloaded from Huggingface to obtain caption embeddings using a pre-trained transformer. However, when we clustered the data in this embedding space, qualitative inspection revealed that the resulting groups lacked coherent driving characteristics. Hence, we experimented with clustering on the TF-IDF features which produced much more coherent clusters with directly interpretable feature space.

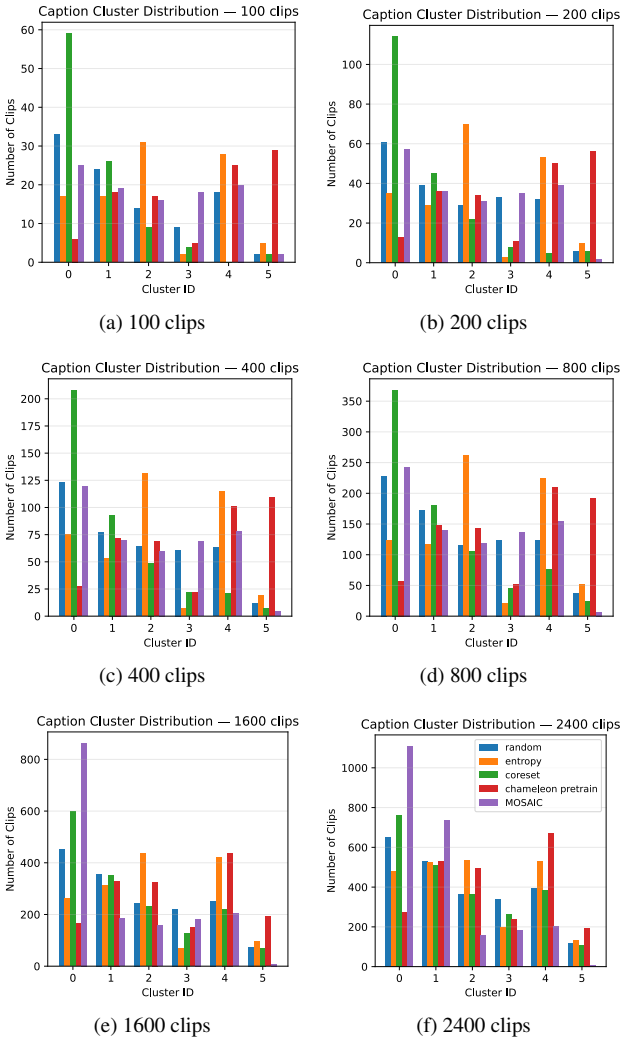


(a) 100 clips (b) 200 clips

(c) 400 clips (d) 800 clips

(e) 1600 clips (f) 2400 clips

Figure 9. Overall caption describing all six subfigures.



(a) 100 clips (b) 200 clips

(c) 400 clips (d) 800 clips

(e) 1600 clips (f) 2400 clips

Figure 10. Overall caption describing all six subfigures.

9. Details on the Scaling Fits and Compute Budget.

MOSAIC requires an upfront compute investment to estimate cluster-specific scaling curves via pilot runs. To keep this cost tractable, we avoid full training from-scratch during the pilot experiments. Instead, we adopt a continual-training approach: we resume training from the base model's final epoch checkpoint and fine-tune on the combined dataset for a small number of epochs. For the OpenScene experiments, we train for 5 epochs after mining 200 and 400 clips from each cluster. For the Navtrain experiments, we train for 10 epochs after mining 100 and 200 clips in the two pilot runs. This procedure provides accurate scaling estimates while maintaining a manageable computational overhead.

For the OpenScene experiments, we share the results

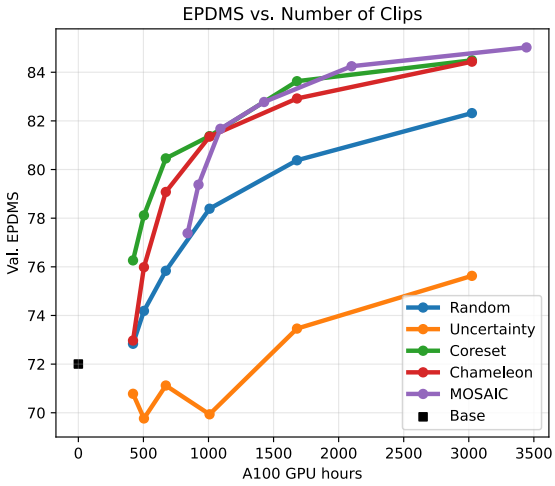


Figure 11. Validation EPDMS vs. Compute Spent (GPU hours) for OpenScene experiments.

with respect to the compute spent for each method. In particular, we provide the validation EPDMS vs. A100 GPU hours. The results are shared in Figure 11. As can be seen while MOSAIC is not the strongest method at small compute budgets, its initial scaling overhead amortizes over time, and at large budgets, the investment in scaling pays off, making MOSAIC the top-performing approach. More concretely, at the highest compute budget: MOSAIC reaches the top-baseline(Coreset in this setting) performance with 16% less compute, corresponding to 490 GPU hours saved; Compared to Random selection, MOSAIC requires 57% less compute, saving 1700 GPU hours to attain the same EPDMS. These results demonstrate that although MOSAIC pays an upfront cost for pilot scaling runs, the compute investment is recovered once we move into the large-budget regime.

10. Ranking with Alternative Cheap Signals

Since ranking is one of the key components of our framework, we also investigate cheaper alternatives to the EPDMS-based ranking signal to reduce the reliance on dense annotations such as bounding boxes. Specifically, we experiment with ranking clips according to (i) the trajectory imitation loss, (ii) the norm of the gradient vector induced by this loss, and (iii) the sensitivity of the model's output to gradient perturbations.

Instead of retraining the model with clips selected using the alternative signals and reporting the validation EPDMS, we measure the Kendall–Tau correlation coefficient between the rankings produced by each alternative signal and those produced by the EPDMS-based ranking. The results, shown in Figure 12, indicate that none of the inexpensive alternatives yield a ranking that correlates strongly with EPDMS.

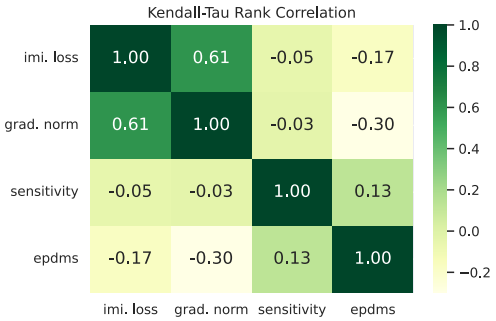


Figure 12. Kendall-Tau correlation coefficients between EPDMS and cheap signals based rankings.

11. Approximation for Linear Separability and Error Analysis:

Here, we formally express the performance improvement obtained from a data mixture $\Delta U(n_1, \cdots, n_M)$ as follows:

$$\sum_{i=1}^{M} \Delta U_i(n_i) + \sum_{i \neq j} \Delta U_{ij}(n_i, n_j) + \text{ H.O.T.}$$

Here, the pairwise cross-cluster interaction term $\Delta U_{ij}(n_i, n_j)$ is defined as $\Delta U_{ij} = U_{ij} - U_i - U_j + U_0$, where we use a lightweight notation for clarity: $U_{ij} = U(\mathcal{D}_{train} \cup \mathcal{D}^i_{sel} \cup \mathcal{D}^j_{sel})$, $U_i = U(\mathcal{D}_{train} \cup \mathcal{D}^i_{sel})$, and $U_0 = U(\mathcal{D}_{train})$, with $U(\cdot) \equiv U(\{\mathcal{G}_r(\cdot)\}_{r=1}^{R})$. In Equation 3, we retain only the first-order terms $\{\Delta U_i\}_{i=1}^{M}$ and omit interaction and higher-order terms. Importantly, we do not assume strict linear separability. Rather, we assume that first-order cluster-wise scaling captures the dominant variation in performance, while interaction terms contribute residual approximation error.

To quantify the magnitude of the approximation error, we compare the *estimated* EPDMS calculated by summing cluster-wise scaling fits against the *actual* EPDMS obtained with the MOSAIC data mixtures. As shown in Table 5, the approximation overestimates performance by a modest margin (up to 1 EPMS), indicating that interaction terms are present but negligible in this setting.

Table 5. Actual vs. estimated EPDMS (Navtrain, geolocation)

# Clips	100	200	400	800	1600	2400
Actual	86.3	87.1	88.2	89.1	90.2	90.3
Estimated	86.2	87.6	89.3	90.6	91.1	91.3

We also note that the discrepancy between the *Actual* and *Estimated* are accumulation of two factors: i) the cross-cluster interactions, ii) extrapolation errors of the scaling fits. Hence, Table 5 should be interpreted as an upper bound on interaction effects rather than a pure estimate thereof.

Also, as a contrasting example, if clusters were formed randomly and lacked semantic coherence, cross-cluster interactions would likely be large, and the approximation would break down. In such pathological settings, explicitly modeling interaction terms would be necessary for optimal data selection.

Table 6. Openscene validation EPDM and BRMR results.

Budget	Method	EPDMS	SRR
250	Random	$72.84_{\pm 1.14}$	1.00
	Uncertainty	$70.78_{\pm 0.59}$	14.58
	Coreset	$76.26_{\pm 0.48}$	0.20
	Chameleon	$72.97_{\pm 1.72}$	0.86
	MOSAIC	$77.38_{\pm 1.58}$	0.15
500	Random	$74.19_{\pm 1.05}$	1.00
	Uncertainty	$69.77_{\pm 0.48}$	10.68
	Coreset	$78.12_{\pm 0.87}$	0.26
	Chameleon	$75.98_{\pm 0.06}$	0.70
	MOSAIC	$79.38_{\pm 1.05}$	0.20
1000	Random	$75.84_{\pm 0.9}$	1.00
	Uncertainty	$71.12_{\pm 0.38}$	NA
	Coreset	$80.46_{\pm 0.02}$	0.28
	Chameleon	$79.08_{\pm 0.74}$	0.44
	MOSAIC	$81.68_{\pm 0.52}$	0.19
2000	Random	$78.39_{\pm 0.12}$	1.00
	Uncertainty	$69.94_{\pm 1.4}$	NA
	Coreset	$81.37_{\pm 0.13}$	0.28
	Chameleon	$81.35_{\pm 0.39}$	0.44
	MOSAIC	$82.78_{\pm 0.41}$	0.19
4000	Random	$80.38_{\pm 0.55}$	1.00
	Uncertainty	$73.46_{\pm 0.19}$	NA
	Coreset	$83.63_{\pm 0.36}$	0.25
	Chameleon	$82.92_{\pm 0.13}$	0.39
	MOSAIC	$84.25_{\pm 0.14}$	0.18
8000	Random	$82.32_{\pm 0.54}$	1.00
	Uncertainty	$75.63_{\pm 0.19}$	NA
	Coreset	$84.49_{\pm 0.02}$	0.35
	Chameleon	$84.43_{\pm 0.01}$	0.40
	MOSAIC	$85.02_{\pm 0.18}$	0.20

Table 7. Navtrain validation EPDMS and BRMR results.

Budget	Method	EPDMS	SRR
100	Random	$84.66_{\pm 0.6}$	1.00
	Uncertainty	$84.5_{\pm 0.48}$	1.47
	Coreset	$85.29_{\pm 0.47}$	0.53
	Chameleon	$84.57_{\pm 0.18}$	1.07
	MOSAIC	$86.29_{\pm 0.43}$	0.30
200	Random	$85.45_{\pm 0.09}$	1.00
	Uncertainty	$84.84_{\pm 0.54}$	1.50
	Coreset	$86.12_{\pm 0.31}$	0.60
	Chameleon	$86.04_{\pm 0.3}$	0.80
	MOSAIC	$87.04_{\pm 0.37}$	0.32
400	Random	$86.69_{\pm 0.2}$	1.00
	Uncertainty	$86.07_{\pm 0.75}$	2.00
	Coreset	$87.09_{\pm 0.29}$	0.79
	Chameleon	$87.04_{\pm 0.6}$	0.82
	MOSAIC	$88.21_{\pm 0.03}$	0.38
800	Random	$87.41_{\pm 0.37}$	1.00
	Uncertainty	$86.69_{\pm 0.34}$	1.69
	Coreset	$88.48_{\pm 0.12}$	0.62
	Chameleon	$88.33_{\pm 0.23}$	0.64
	MOSAIC	$89.1_{\pm 0.12}$	0.33
1600	Random	$88.62_{\pm 0.22}$	1.00
	Uncertainty	$87.75_{\pm 0.37}$	1.36
	Coreset	$89.3_{\pm 0.19}$	0.58
	Chameleon	$89.5_{\pm 0.2}$	0.62
	MOSAIC	$90.18_{\pm 0.25}$	0.37
2400	Random	$89.42_{\pm 0.03}$	1.00
	Uncertainty	$88.95_{\pm 0.15}$	1.00
	Coreset	$89.75_{\pm 0.02}$	0.76
	Chameleon	$90.05_{\pm 0.08}$	0.64
	MOSAIC	$90.31_{\pm 0.03}$	0.43

Table 8. Navtrain validation EPDMS and BRMR results under caption-based clustering.

Budget	Method	EPDMS	SRR
100	Random	$84.66_{\pm 0.6}$	1.00
	Uncertainty	$84.5_{\pm 0.48}$	1.47
	Coreset	$85.29_{\pm 0.47}$	0.53
	Chameleon	$84.35_{\pm 0.47}$	1.30
	MOSAIC	$85.85_{\pm 0.41}$	0.37
200	Random	$85.45_{\pm 0.09}$	1.00
	Uncertainty	$84.84_{\pm 0.54}$	1.50
	Coreset	$86.12_{\pm 0.31}$	0.60
	Chameleon	$85.39_{\pm 0.02}$	2.88
	MOSAIC	$86.75_{\pm 0.17}$	0.40
400	Random	$86.69_{\pm 0.2}$	1.00
	Uncertainty	$86.07_{\pm 0.75}$	2.00
	Coreset	$87.09_{\pm 0.29}$	0.79
	Chameleon	$84.95_{\pm 0.45}$	3.32
	MOSAIC	$88.11_{\pm 0.05}$	0.48
800	Random	$87.41_{\pm 0.37}$	1.00
	Uncertainty	$86.69_{\pm 0.34}$	1.69
	Coreset	$88.48_{\pm 0.12}$	0.62
	Chameleon	$86.1_{\pm 0.55}$	2.68
	MOSAIC	$88.99_{\pm 0.09}$	0.37
1600	Random	$88.62_{\pm 0.22}$	1.00
	Uncertainty	$87.75_{\pm 0.37}$	1.36
	Coreset	$89.3_{\pm 0.19}$	0.58
	Chameleon	$86.99_{\pm 0.57}$	1.50
	MOSAIC	$89.98_{\pm 0.13}$	0.39
2400	Random	$89.42_{\pm 0.03}$	1.00
	Uncertainty	$88.95_{\pm 0.15}$	1.00
	Coreset	$89.75_{\pm 0.02}$	0.76
	Chameleon	$87.62_{\pm 0.28}$	1.00
	MOSAIC	$90.37_{\pm 0.2}$	0.48

Table 9. Breakdown of the nine EPDMS rule-compliance metrics for the base model and the models trained with data selected by various strategies at all budgets, shown for the OpenScene experiment.

Setting		NC	DAC	DDC	TLC	EP	TTC	LK	HC	EC	EPDMS
Base		94.05	83.9	96.28	99.6	85.96	92.95	93.26	98.25	81.88	72.0
250	Random	$94.27_{\pm 0.60}$	$84.63_{\pm 1.46}$	$97.38_{\pm 0.23}$	$99.66_{\pm 0.04}$	$85.18_{\pm 1.02}$	$93.23_{\pm 0.64}$	$93.33_{\pm 0.56}$	$98.26_{\pm 0.01}$	$82.66_{\pm 0.76}$	$72.84_{\pm 1.14}$
	Uncertainty	$93.97_{\pm 0.44}$	$82.49_{\pm 0.30}$	$96.78_{\pm 0.44}$	$99.66_{\pm 0.02}$	$85.18_{\pm 0.81}$	$92.98_{\pm 0.42}$	$93.18_{\pm 0.66}$	$98.23_{\pm 0.08}$	$82.15_{\pm 0.27}$	$70.78_{\pm 0.59}$
	Coreset	$95.11_{\pm 0.47}$	$87.66_{\pm 0.61}$	$98.38_{\pm 0.21}$	$99.67_{\pm 0.04}$	$86.09_{\pm 1.13}$	$94.08_{\pm 0.84}$	$94.47_{\pm 0.20}$	$98.31_{\pm 0.05}$	$83.38_{\pm 0.74}$	$76.26_{\pm 0.48}$
	Chameleon	$94.02_{\pm 1.25}$	$84.30_{\pm 1.18}$	$97.48_{\pm 0.71}$	$99.58_{\pm 0.06}$	$87.48_{\pm 1.41}$	$92.69_{\pm 1.23}$	$93.43_{\pm 0.04}$	$98.26_{\pm 0.01}$	$83.15_{\pm 1.80}$	$72.97_{\pm 1.72}$
	MOSAIC	$94.89_{\pm 0.74}$	$88.76_{\pm 1.17}$	$98.54_{\pm 0.43}$	$99.61_{\pm 0.04}$	$86.50_{\pm 1.03}$	$93.93_{\pm 0.88}$	$94.88_{\pm 0.14}$	$98.26_{\pm 0.03}$	$83.77_{\pm 0.67}$	$77.38_{\pm 1.58}$
500	Random	$94.65_{\pm 0.21}$	$85.72_{\pm 0.88}$	$97.87_{\pm 0.44}$	$99.64_{\pm 0.06}$	$85.53_{\pm 0.22}$	$93.51_{\pm 0.32}$	$93.73_{\pm 0.24}$	$98.27_{\pm 0.05}$	$83.26_{\pm 0.14}$	$74.19_{\pm 1.05}$
	Uncertainty	$93.32_{\pm 0.47}$	$82.26_{\pm 0.40}$	$96.09_{\pm 0.52}$	$99.60_{\pm 0.08}$	$84.51_{\pm 0.43}$	$92.23_{\pm 0.73}$	$92.38_{\pm 0.56}$	$98.30_{\pm 0.01}$	$82.85_{\pm 1.07}$	$69.77_{\pm 0.48}$
	Coreset	$95.56_{\pm 0.78}$	$88.96_{\pm 0.57}$	$98.95_{\pm 0.09}$	$99.71_{\pm 0.07}$	$86.21_{\pm 0.99}$	$94.69_{\pm 0.79}$	$95.14_{\pm 0.13}$	$98.31_{\pm 0.03}$	$84.24_{\pm 0.34}$	$78.12_{\pm 0.87}$
	Chameleon	$95.00_{\pm 0.58}$	$87.11_{\pm 0.09}$	$98.16_{\pm 0.02}$	$99.67_{\pm 0.16}$	$86.67_{\pm 2.25}$	$94.22_{\pm 0.45}$	$94.20_{\pm 0.44}$	$98.30_{\pm 0.01}$	$83.69_{\pm 0.24}$	$75.98_{\pm 0.06}$
	MOSAIC	$95.57_{\pm 1.05}$	$90.54_{\pm 0.45}$	$98.83_{\pm 0.29}$	$99.67_{\pm 0.09}$	$86.08_{\pm 1.79}$	$94.85_{\pm 1.23}$	$95.68_{\pm 0.27}$	$98.25_{\pm 0.04}$	$83.80_{\pm 0.16}$	$79.38_{\pm 1.05}$
1000	Random	$95.21_{\pm 0.58}$	$87.15_{\pm 1.44}$	$98.26_{\pm 0.39}$	$99.72_{\pm 0.07}$	$85.56_{\pm 0.96}$	$94.35_{\pm 0.60}$	$94.50_{\pm 0.66}$	$98.31_{\pm 0.03}$	$82.50_{\pm 0.52}$	$75.84_{\pm 0.90}$
	Uncertainty	$94.04_{\pm 0.70}$	$83.77_{\pm 0.02}$	$96.96_{\pm 0.08}$	$99.70_{\pm 0.08}$	$83.11_{\pm 0.66}$	$93.21_{\pm 1.00}$	$92.87_{\pm 0.14}$	$98.32_{\pm 0.02}$	$81.91_{\pm 0.75}$	$71.12_{\pm 0.38}$
	Coreset	$95.93_{\pm 0.24}$	$91.05_{\pm 0.26}$	$99.28_{\pm 0.11}$	$99.71_{\pm 0.04}$	$86.39_{\pm 0.48}$	$95.01_{\pm 0.21}$	$95.75_{\pm 0.08}$	$98.28_{\pm 0.03}$	$84.58_{\pm 0.42}$	$80.46_{\pm 0.02}$
	Chameleon	$95.89_{\pm 0.19}$	$89.57_{\pm 0.68}$	$98.94_{\pm 0.16}$	$99.71_{\pm 0.07}$	$86.39_{\pm 0.51}$	$95.06_{\pm 0.26}$	$95.44_{\pm 0.27}$	$98.29_{\pm 0.01}$	$84.23_{\pm 0.74}$	$79.08_{\pm 0.74}$
	MOSAIC	$96.00_{\pm 0.22}$	$92.20_{\pm 0.48}$	$99.33_{\pm 0.07}$	$99.67_{\pm 0.05}$	$86.63_{\pm 0.41}$	$95.24_{\pm 0.24}$	$96.17_{\pm 0.22}$	$98.28_{\pm 0.03}$	$84.33_{\pm 0.30}$	$81.68_{\pm 0.52}$
2000	Random	$95.58_{\pm 0.54}$	$89.26_{\pm 0.64}$	$98.67_{\pm 0.18}$	$99.70_{\pm 0.12}$	$86.44_{\pm 0.42}$	$94.88_{\pm 0.61}$	$95.26_{\pm 0.18}$	$98.30_{\pm 0.00}$	$83.96_{\pm 0.96}$	$78.39_{\pm 0.12}$
	Uncertainty	$93.14_{\pm 0.79}$	$82.66_{\pm 1.12}$	$96.64_{\pm 0.64}$	$99.53_{\pm 0.09}$	$84.52_{\pm 1.23}$	$92.19_{\pm 1.22}$	$93.22_{\pm 0.29}$	$98.28_{\pm 0.03}$	$80.98_{\pm 1.30}$	$69.94_{\pm 1.40}$
	Coreset	$95.89_{\pm 0.22}$	$91.77_{\pm 0.14}$	$99.44_{\pm 0.06}$	$99.66_{\pm 0.04}$	$87.39_{\pm 0.05}$	$94.98_{\pm 0.18}$	$95.99_{\pm 0.47}$	$98.29_{\pm 0.00}$	$85.55_{\pm 0.19}$	$81.37_{\pm 0.13}$
	Chameleon	$96.38_{\pm 0.25}$	$91.31_{\pm 0.26}$	$99.15_{\pm 0.03}$	$99.71_{\pm 0.05}$	$86.55_{\pm 0.40}$	$95.60_{\pm 0.29}$	$95.99_{\pm 0.15}$	$98.34_{\pm 0.01}$	$85.04_{\pm 0.15}$	$81.35_{\pm 0.39}$
	MOSAIC	$96.90_{\pm 0.38}$	$92.29_{\pm 0.36}$	$99.48_{\pm 0.05}$	$99.73_{\pm 0.01}$	$86.61_{\pm 0.73}$	$96.16_{\pm 0.26}$	$96.34_{\pm 0.06}$	$98.28_{\pm 0.05}$	$84.69_{\pm 0.01}$	$82.78_{\pm 0.41}$
4000	Random	$96.32_{\pm 0.59}$	$90.53_{\pm 0.06}$	$99.06_{\pm 0.07}$	$99.79_{\pm 0.05}$	$86.36_{\pm 0.48}$	$95.66_{\pm 0.52}$	$95.68_{\pm 0.09}$	$98.30_{\pm 0.01}$	$84.46_{\pm 0.14}$	$80.38_{\pm 0.55}$
	Uncertainty	$94.67_{\pm 0.28}$	$85.11_{\pm 0.51}$	$97.15_{\pm 0.54}$	$99.71_{\pm 0.04}$	$84.26_{\pm 0.69}$	$93.72_{\pm 0.40}$	$93.26_{\pm 0.09}$	$98.28_{\pm 0.02}$	$81.34_{\pm 1.06}$	$73.46_{\pm 0.19}$
	Coreset	$97.11_{\pm 0.18}$	$92.93_{\pm 0.60}$	$99.44_{\pm 0.06}$	$99.82_{\pm 0.02}$	$86.65_{\pm 0.55}$	$96.42_{\pm 0.19}$	$96.66_{\pm 0.30}$	$98.16_{\pm 0.12}$	$85.10_{\pm 0.06}$	$83.63_{\pm 0.36}$
	Chameleon	$96.76_{\pm 0.24}$	$92.32_{\pm 0.02}$	$99.51_{\pm 0.01}$	$99.77_{\pm 0.01}$	$86.98_{\pm 0.17}$	$95.91_{\pm 0.31}$	$96.49_{\pm 0.12}$	$98.32_{\pm 0.01}$	$85.51_{\pm 0.11}$	$82.92_{\pm 0.13}$
	MOSAIC	$96.97_{\pm 0.32}$	$93.59_{\pm 0.11}$	$99.59_{\pm 0.04}$	$99.80_{\pm 0.01}$	$87.14_{\pm 0.98}$	$96.18_{\pm 0.45}$	$96.62_{\pm 0.08}$	$98.28_{\pm 0.01}$	$85.06_{\pm 0.34}$	$84.25_{\pm 0.14}$
8000	Random	$96.79_{\pm 0.21}$	$91.88_{\pm 0.34}$	$99.23_{\pm 0.11}$	$99.79_{\pm 0.03}$	$87.19_{\pm 0.05}$	$95.93_{\pm 0.15}$	$96.19_{\pm 0.10}$	$98.28_{\pm 0.03}$	$84.97_{\pm 0.19}$	$82.32_{\pm 0.54}$
	Uncertainty	$95.62_{\pm 0.38}$	$86.48_{\pm 0.06}$	$97.62_{\pm 0.01}$	$99.71_{\pm 0.02}$	$84.92_{\pm 0.25}$	$94.80_{\pm 0.28}$	$94.34_{\pm 0.27}$	$98.32_{\pm 0.02}$	$81.62_{\pm 0.09}$	$75.63_{\pm 0.19}$
	Coreset	$97.39_{\pm 0.15}$	$93.51_{\pm 0.18}$	$99.55_{\pm 0.07}$	$99.81_{\pm 0.03}$	$87.07_{\pm 0.39}$	$96.64_{\pm 0.12}$	$96.78_{\pm 0.06}$	$98.28_{\pm 0.03}$	$85.51_{\pm 0.15}$	$84.49_{\pm 0.02}$
	Chameleon	$97.33_{\pm 0.39}$	$93.36_{\pm 0.14}$	$99.61_{\pm 0.01}$	$99.82_{\pm 0.01}$	$87.34_{\pm 0.61}$	$96.42_{\pm 0.50}$	$96.90_{\pm 0.17}$	$98.29_{\pm 0.02}$	$85.51_{\pm 0.12}$	$84.43_{\pm 0.00}$
	MOSAIC	$97.55_{\pm 0.13}$	$93.84_{\pm 0.00}$	$99.53_{\pm 0.18}$	$99.84_{\pm 0.03}$	$87.19_{\pm 0.24}$	$96.79_{\pm 0.07}$	$97.10_{\pm 0.07}$	$98.29_{\pm 0.02}$	$85.25_{\pm 0.22}$	$85.02_{\pm 0.18}$

Table 10. Breakdown of the nine EPDMS rule-compliance metrics for the base model and the models trained with data selected by various strategies at all budgets, shown for the Navtrain experiment.

Setting		NC	DAC	DDC	TLC	EP	TTC	LK	HC	EC	EPDMS
Base		95.3	95.94	99.09	99.6	88.09	94.55	94.49	98.25	82.39	83.97
100	Random	$95.43_{\pm 0.84}$	$96.41_{\pm 0.20}$	$98.98_{\pm 0.07}$	$99.54_{\pm 0.15}$	$88.68_{\pm 0.63}$	$94.69_{\pm 0.89}$	$94.82_{\pm 0.34}$	$98.27_{\pm 0.04}$	$82.81_{\pm 0.64}$	$84.66_{\pm 0.60}$
	Uncertainty	$95.68_{\pm 0.33}$	$96.23_{\pm 0.38}$	$98.91_{\pm 0.12}$	$99.51_{\pm 0.06}$	$88.21_{\pm 0.22}$	$94.77_{\pm 0.36}$	$94.88_{\pm 0.13}$	$98.27_{\pm 0.04}$	$83.50_{\pm 0.32}$	$84.50_{\pm 0.48}$
	Coreset	$95.63_{\pm 0.41}$	$96.88_{\pm 0.33}$	$99.13_{\pm 0.09}$	$99.56_{\pm 0.03}$	$88.39_{\pm 0.65}$	$94.75_{\pm 0.51}$	$94.97_{\pm 0.34}$	$98.25_{\pm 0.03}$	$82.94_{\pm 0.20}$	$85.29_{\pm 0.47}$
	Chameleon	$95.14_{\pm 0.20}$	$96.50_{\pm 0.19}$	$99.17_{\pm 0.02}$	$99.53_{\pm 0.02}$	$88.80_{\pm 0.18}$	$94.35_{\pm 0.11}$	$95.10_{\pm 0.12}$	$98.25_{\pm 0.04}$	$82.93_{\pm 0.75}$	$84.57_{\pm 0.18}$
	MOSAIC	$96.75_{\pm 0.28}$	$97.06_{\pm 0.09}$	$99.03_{\pm 0.03}$	$99.60_{\pm 0.03}$	$87.74_{\pm 0.28}$	$96.09_{\pm 0.35}$	$94.92_{\pm 0.32}$	$98.27_{\pm 0.02}$	$82.80_{\pm 0.62}$	$86.29_{\pm 0.43}$
200	Random	$95.90_{\pm 0.35}$	$96.58_{\pm 0.23}$	$99.11_{\pm 0.08}$	$99.65_{\pm 0.02}$	$88.75_{\pm 0.19}$	$95.08_{\pm 0.33}$	$95.14_{\pm 0.29}$	$98.27_{\pm 0.04}$	$83.14_{\pm 0.11}$	$85.45_{\pm 0.09}$
	Uncertainty	$95.61_{\pm 0.66}$	$96.53_{\pm 0.38}$	$98.96_{\pm 0.18}$	$99.57_{\pm 0.09}$	$88.51_{\pm 0.11}$	$94.74_{\pm 0.59}$	$94.88_{\pm 0.28}$	$98.28_{\pm 0.03}$	$83.34_{\pm 0.34}$	$84.84_{\pm 0.54}$
	Coreset	$96.19_{\pm 0.49}$	$97.05_{\pm 0.11}$	$99.13_{\pm 0.06}$	$99.60_{\pm 0.04}$	$88.68_{\pm 0.13}$	$95.39_{\pm 0.44}$	$95.17_{\pm 0.11}$	$98.29_{\pm 0.01}$	$83.45_{\pm 0.55}$	$86.12_{\pm 0.31}$
	Chameleon	$96.12_{\pm 0.52}$	$96.76_{\pm 0.34}$	$99.33_{\pm 0.18}$	$99.60_{\pm 0.10}$	$88.74_{\pm 0.54}$	$95.38_{\pm 0.71}$	$95.41_{\pm 0.19}$	$98.30_{\pm 0.02}$	$83.73_{\pm 0.13}$	$86.04_{\pm 0.30}$
	MOSAIC	$96.83_{\pm 0.31}$	$97.51_{\pm 0.18}$	$99.24_{\pm 0.06}$	$99.61_{\pm 0.01}$	$88.20_{\pm 0.13}$	$96.16_{\pm 0.29}$	$95.36_{\pm 0.18}$	$98.26_{\pm 0.02}$	$82.63_{\pm 0.35}$	$87.04_{\pm 0.37}$
400	Random	$96.71_{\pm 0.25}$	$96.91_{\pm 0.20}$	$99.18_{\pm 0.09}$	$99.71_{\pm 0.01}$	$88.75_{\pm 0.15}$	$96.02_{\pm 0.23}$	$95.76_{\pm 0.16}$	$98.30_{\pm 0.01}$	$82.96_{\pm 0.10}$	$86.69_{\pm 0.20}$
	Uncertainty	$96.39_{\pm 0.60}$	$96.97_{\pm 0.38}$	$99.00_{\pm 0.08}$	$99.65_{\pm 0.01}$	$88.22_{\pm 0.42}$	$95.55_{\pm 0.66}$	$94.98_{\pm 0.22}$	$98.25_{\pm 0.02}$	$83.64_{\pm 0.16}$	$86.07_{\pm 0.75}$
	Coreset	$96.73_{\pm 0.24}$	$97.27_{\pm 0.17}$	$99.36_{\pm 0.02}$	$99.64_{\pm 0.02}$	$88.80_{\pm 0.11}$	$95.95_{\pm 0.26}$	$95.81_{\pm 0.23}$	$98.29_{\pm 0.03}$	$83.48_{\pm 0.46}$	$87.09_{\pm 0.29}$
	Chameleon	$96.33_{\pm 0.36}$	$97.55_{\pm 0.20}$	$99.37_{\pm 0.07}$	$99.63_{\pm 0.03}$	$88.97_{\pm 0.42}$	$95.59_{\pm 0.38}$	$95.87_{\pm 0.20}$	$98.30_{\pm 0.01}$	$83.10_{\pm 0.35}$	$87.04_{\pm 0.60}$
	MOSAIC	$97.75_{\pm 0.08}$	$97.79_{\pm 0.11}$	$99.42_{\pm 0.06}$	$99.72_{\pm 0.04}$	$87.62_{\pm 0.11}$	$97.17_{\pm 0.09}$	$95.54_{\pm 0.08}$	$98.24_{\pm 0.01}$	$82.81_{\pm 0.27}$	$88.21_{\pm 0.03}$
800	Random	$96.94_{\pm 0.35}$	$97.15_{\pm 0.36}$	$99.35_{\pm 0.12}$	$99.69_{\pm 0.05}$	$89.16_{\pm 0.06}$	$96.22_{\pm 0.41}$	$96.28_{\pm 0.45}$	$98.29_{\pm 0.03}$	$83.63_{\pm 0.02}$	$87.41_{\pm 0.37}$
	Uncertainty	$96.98_{\pm 0.40}$	$96.88_{\pm 0.16}$	$99.13_{\pm 0.11}$	$99.69_{\pm 0.07}$	$88.31_{\pm 0.45}$	$96.22_{\pm 0.32}$	$95.42_{\pm 0.15}$	$98.28_{\pm 0.03}$	$82.95_{\pm 0.31}$	$86.69_{\pm 0.34}$
	Coreset	$97.21_{\pm 0.12}$	$98.06_{\pm 0.23}$	$99.49_{\pm 0.06}$	$99.67_{\pm 0.05}$	$88.84_{\pm 0.08}$	$96.62_{\pm 0.13}$	$96.22_{\pm 0.19}$	$98.30_{\pm 0.03}$	$83.67_{\pm 0.27}$	$88.48_{\pm 0.12}$
	Chameleon	$97.07_{\pm 0.17}$	$97.97_{\pm 0.18}$	$99.48_{\pm 0.08}$	$99.68_{\pm 0.03}$	$88.99_{\pm 0.50}$	$96.57_{\pm 0.19}$	$96.38_{\pm 0.18}$	$98.29_{\pm 0.03}$	$83.28_{\pm 0.22}$	$88.33_{\pm 0.23}$
	MOSAIC	$97.65_{\pm 0.14}$	$98.33_{\pm 0.06}$	$99.54_{\pm 0.05}$	$99.73_{\pm 0.05}$	$88.68_{\pm 0.44}$	$97.03_{\pm 0.16}$	$96.19_{\pm 0.14}$	$98.26_{\pm 0.02}$	$82.93_{\pm 0.48}$	$89.10_{\pm 0.12}$
1600	Random	$97.17_{\pm 0.07}$	$98.19_{\pm 0.43}$	$99.42_{\pm 0.05}$	$99.69_{\pm 0.02}$	$89.36_{\pm 0.12}$	$96.50_{\pm 0.14}$	$96.45_{\pm 0.25}$	$98.31_{\pm 0.03}$	$83.17_{\pm 0.76}$	$88.62_{\pm 0.22}$
	Uncertainty	$96.92_{\pm 0.38}$	$97.66_{\pm 0.08}$	$99.22_{\pm 0.10}$	$99.77_{\pm 0.02}$	$89.02_{\pm 0.28}$	$96.24_{\pm 0.40}$	$96.10_{\pm 0.07}$	$98.30_{\pm 0.01}$	$82.92_{\pm 0.38}$	$87.75_{\pm 0.37}$
	Coreset	$97.50_{\pm 0.10}$	$98.31_{\pm 0.34}$	$99.59_{\pm 0.03}$	$99.72_{\pm 0.05}$	$89.27_{\pm 0.21}$	$96.86_{\pm 0.07}$	$96.75_{\pm 0.22}$	$98.30_{\pm 0.03}$	$83.88_{\pm 0.50}$	$89.30_{\pm 0.19}$
	Chameleon	$97.43_{\pm 0.22}$	$98.46_{\pm 0.17}$	$99.60_{\pm 0.05}$	$99.75_{\pm 0.03}$	$89.60_{\pm 0.19}$	$96.83_{\pm 0.30}$	$96.89_{\pm 0.07}$	$98.30_{\pm 0.03}$	$83.87_{\pm 0.34}$	$89.50_{\pm 0.20}$
	MOSAIC	$98.04_{\pm 0.24}$	$98.61_{\pm 0.32}$	$99.63_{\pm 0.06}$	$99.73_{\pm 0.02}$	$89.28_{\pm 0.19}$	$97.50_{\pm 0.32}$	$97.07_{\pm 0.06}$	$98.28_{\pm 0.04}$	$83.70_{\pm 0.41}$	$90.18_{\pm 0.25}$
2400	Random	$97.56_{\pm 0.11}$	$98.23_{\pm 0.12}$	$99.56_{\pm 0.04}$	$99.74_{\pm 0.00}$	$89.57_{\pm 0.08}$	$96.97_{\pm 0.09}$	$96.95_{\pm 0.07}$	$98.30_{\pm 0.01}$	$83.95_{\pm 0.31}$	$89.42_{\pm 0.03}$
	Uncertainty	$97.62_{\pm 0.21}$	$98.10_{\pm 0.17}$	$99.36_{\pm 0.10}$	$99.78_{\pm 0.02}$	$89.19_{\pm 0.15}$	$97.07_{\pm 0.22}$	$96.65_{\pm 0.27}$	$98.29_{\pm 0.05}$	$82.53_{\pm 0.48}$	$88.95_{\pm 0.15}$
	Coreset	$97.59_{\pm 0.02}$	$98.53_{\pm 0.06}$	$99.57_{\pm 0.04}$	$99.67_{\pm 0.04}$	$89.79_{\pm 0.24}$	$97.17_{\pm 0.13}$	$97.17_{\pm 0.10}$	$98.31_{\pm 0.02}$	$83.77_{\pm 0.54}$	$89.75_{\pm 0.02}$
	Chameleon	$97.60_{\pm 0.16}$	$98.71_{\pm 0.06}$	$99.63_{\pm 0.04}$	$99.77_{\pm 0.01}$	$89.85_{\pm 0.06}$	$97.18_{\pm 0.14}$	$97.20_{\pm 0.09}$	$98.28_{\pm 0.01}$	$83.61_{\pm 0.51}$	$90.05_{\pm 0.08}$
	MOSAIC	$98.02_{\pm 0.12}$	$98.69_{\pm 0.05}$	$99.66_{\pm 0.07}$	$99.80_{\pm 0.06}$	$89.19_{\pm 0.38}$	$97.58_{\pm 0.10}$	$97.22_{\pm 0.09}$	$98.31_{\pm 0.00}$	$83.56_{\pm 0.07}$	$90.31_{\pm 0.03}$

Table 11. Breakdown of the nine EPDMS rule-compliance metrics for the base model and the models trained with data selected by various strategies at all budgets, shown for the Navtrain experiment when the clustering is performed on the clip captions.

Setting		NC	DAC	DDC	TLC	EP	TTC	LK	HC	EC	EPDMS
Base		95.3	95.94	99.09	99.6	88.09	94.55	94.49	98.25	82.39	83.97
100	Random	$95.43_{\pm 0.84}$	$96.41_{\pm 0.20}$	$98.98_{\pm 0.07}$	$99.54_{\pm 0.15}$	$88.68_{\pm 0.63}$	$94.69_{\pm 0.89}$	$94.82_{\pm 0.34}$	$98.27_{\pm 0.04}$	$82.81_{\pm 0.64}$	$84.66_{\pm 0.60}$
	Uncertainty	$95.68_{\pm 0.33}$	$96.23_{\pm 0.38}$	$98.91_{\pm 0.12}$	$99.51_{\pm 0.06}$	$88.21_{\pm 0.22}$	$94.77_{\pm 0.36}$	$94.88_{\pm 0.13}$	$98.27_{\pm 0.04}$	$83.50_{\pm 0.32}$	$84.50_{\pm 0.48}$
	Coreset	$95.63_{\pm 0.41}$	$96.88_{\pm 0.33}$	$99.13_{\pm 0.09}$	$99.56_{\pm 0.03}$	$88.39_{\pm 0.65}$	$94.75_{\pm 0.51}$	$94.97_{\pm 0.34}$	$98.25_{\pm 0.03}$	$82.94_{\pm 0.20}$	$85.29_{\pm 0.47}$
	Chameleon	$95.43_{\pm 0.61}$	$96.14_{\pm 0.02}$	$98.94_{\pm 0.12}$	$99.56_{\pm 0.06}$	$88.45_{\pm 0.26}$	$94.52_{\pm 0.59}$	$94.82_{\pm 0.07}$	$98.28_{\pm 0.02}$	$83.27_{\pm 0.49}$	$84.35_{\pm 0.47}$
	MOSAIC	$96.53_{\pm 0.31}$	$96.91_{\pm 0.30}$	$99.03_{\pm 0.12}$	$99.54_{\pm 0.06}$	$87.62_{\pm 0.21}$	$95.80_{\pm 0.32}$	$94.85_{\pm 0.34}$	$98.24_{\pm 0.02}$	$82.66_{\pm 0.66}$	$85.85_{\pm 0.41}$
200	Random	$95.90_{\pm 0.35}$	$96.58_{\pm 0.23}$	$99.11_{\pm 0.08}$	$99.65_{\pm 0.02}$	$88.75_{\pm 0.19}$	$95.08_{\pm 0.33}$	$95.14_{\pm 0.29}$	$98.27_{\pm 0.04}$	$83.14_{\pm 0.11}$	$85.45_{\pm 0.09}$
	Uncertainty	$95.61_{\pm 0.66}$	$96.53_{\pm 0.38}$	$98.96_{\pm 0.18}$	$99.57_{\pm 0.09}$	$88.51_{\pm 0.11}$	$94.74_{\pm 0.59}$	$94.88_{\pm 0.28}$	$98.28_{\pm 0.03}$	$83.34_{\pm 0.34}$	$84.84_{\pm 0.54}$
	Coreset	$96.19_{\pm 0.49}$	$97.05_{\pm 0.11}$	$99.13_{\pm 0.06}$	$99.60_{\pm 0.04}$	$88.68_{\pm 0.13}$	$95.39_{\pm 0.44}$	$95.17_{\pm 0.11}$	$98.29_{\pm 0.01}$	$83.45_{\pm 0.55}$	$86.12_{\pm 0.31}$
	Chameleon	$96.20_{\pm 0.11}$	$96.58_{\pm 0.25}$	$98.97_{\pm 0.21}$	$99.63_{\pm 0.01}$	$88.02_{\pm 0.19}$	$95.42_{\pm 0.24}$	$94.88_{\pm 0.23}$	$98.30_{\pm 0.00}$	$82.88_{\pm 1.50}$	$85.39_{\pm 0.02}$
	MOSAIC	$97.07_{\pm 0.30}$	$97.19_{\pm 0.25}$	$99.08_{\pm 0.06}$	$99.64_{\pm 0.03}$	$87.79_{\pm 0.46}$	$96.28_{\pm 0.36}$	$94.99_{\pm 0.29}$	$98.25_{\pm 0.01}$	$82.92_{\pm 0.55}$	$86.75_{\pm 0.17}$
400	Random	$96.71_{\pm 0.25}$	$96.91_{\pm 0.20}$	$99.18_{\pm 0.09}$	$99.71_{\pm 0.01}$	$88.75_{\pm 0.15}$	$96.02_{\pm 0.23}$	$95.76_{\pm 0.16}$	$98.30_{\pm 0.01}$	$82.96_{\pm 0.10}$	$86.69_{\pm 0.20}$
	Uncertainty	$96.39_{\pm 0.60}$	$96.97_{\pm 0.38}$	$99.00_{\pm 0.08}$	$99.65_{\pm 0.01}$	$88.22_{\pm 0.42}$	$95.55_{\pm 0.66}$	$94.98_{\pm 0.22}$	$98.25_{\pm 0.02}$	$83.64_{\pm 0.16}$	$86.07_{\pm 0.75}$
	Coreset	$96.73_{\pm 0.24}$	$97.27_{\pm 0.17}$	$99.36_{\pm 0.02}$	$99.64_{\pm 0.02}$	$88.80_{\pm 0.11}$	$95.95_{\pm 0.26}$	$95.81_{\pm 0.23}$	$98.29_{\pm 0.03}$	$83.48_{\pm 0.46}$	$87.09_{\pm 0.29}$
	Chameleon	$95.61_{\pm 0.32}$	$96.50_{\pm 0.25}$	$99.04_{\pm 0.11}$	$99.59_{\pm 0.06}$	$88.64_{\pm 0.56}$	$94.84_{\pm 0.42}$	$94.90_{\pm 0.13}$	$98.29_{\pm 0.00}$	$82.73_{\pm 0.34}$	$84.95_{\pm 0.45}$
	MOSAIC	$97.36_{\pm 0.10}$	$97.91_{\pm 0.05}$	$99.33_{\pm 0.10}$	$99.66_{\pm 0.02}$	$88.37_{\pm 0.41}$	$96.68_{\pm 0.11}$	$95.43_{\pm 0.34}$	$98.27_{\pm 0.03}$	$83.00_{\pm 1.38}$	$88.11_{\pm 0.05}$
800	Random	$96.94_{\pm 0.35}$	$97.15_{\pm 0.36}$	$99.35_{\pm 0.12}$	$99.69_{\pm 0.05}$	$89.16_{\pm 0.06}$	$96.22_{\pm 0.41}$	$96.28_{\pm 0.45}$	$98.29_{\pm 0.03}$	$83.63_{\pm 0.02}$	$87.41_{\pm 0.37}$
	Uncertainty	$96.98_{\pm 0.40}$	$96.88_{\pm 0.16}$	$99.13_{\pm 0.11}$	$99.69_{\pm 0.07}$	$88.31_{\pm 0.45}$	$96.22_{\pm 0.32}$	$95.42_{\pm 0.15}$	$98.28_{\pm 0.03}$	$82.95_{\pm 0.31}$	$86.69_{\pm 0.34}$
	Coreset	$97.21_{\pm 0.12}$	$98.06_{\pm 0.23}$	$99.49_{\pm 0.06}$	$99.67_{\pm 0.05}$	$88.84_{\pm 0.08}$	$96.62_{\pm 0.13}$	$96.22_{\pm 0.19}$	$98.30_{\pm 0.03}$	$83.67_{\pm 0.27}$	$88.48_{\pm 0.12}$
	Chameleon	$96.26_{\pm 0.49}$	$96.83_{\pm 0.47}$	$99.10_{\pm 0.05}$	$99.71_{\pm 0.03}$	$88.89_{\pm 0.48}$	$95.53_{\pm 0.46}$	$95.64_{\pm 0.21}$	$98.28_{\pm 0.01}$	$82.97_{\pm 0.19}$	$86.10_{\pm 0.55}$
	MOSAIC	$97.92_{\pm 0.09}$	$98.08_{\pm 0.17}$	$99.50_{\pm 0.05}$	$99.73_{\pm 0.01}$	$88.20_{\pm 0.25}$	$97.35_{\pm 0.14}$	$96.12_{\pm 0.22}$	$98.25_{\pm 0.04}$	$83.00_{\pm 0.50}$	$88.99_{\pm 0.09}$
1600	Random	$97.17_{\pm 0.07}$	$98.19_{\pm 0.43}$	$99.42_{\pm 0.05}$	$99.69_{\pm 0.02}$	$89.36_{\pm 0.12}$	$96.50_{\pm 0.14}$	$96.45_{\pm 0.25}$	$98.31_{\pm 0.03}$	$83.17_{\pm 0.76}$	$88.62_{\pm 0.22}$
	Uncertainty	$96.92_{\pm 0.38}$	$97.66_{\pm 0.08}$	$99.22_{\pm 0.10}$	$99.77_{\pm 0.02}$	$89.02_{\pm 0.28}$	$96.24_{\pm 0.40}$	$96.10_{\pm 0.07}$	$98.30_{\pm 0.01}$	$82.92_{\pm 0.38}$	$87.75_{\pm 0.37}$
	Coreset	$97.50_{\pm 0.10}$	$98.31_{\pm 0.34}$	$99.59_{\pm 0.03}$	$99.72_{\pm 0.05}$	$89.27_{\pm 0.21}$	$96.86_{\pm 0.07}$	$96.75_{\pm 0.22}$	$98.30_{\pm 0.03}$	$83.88_{\pm 0.50}$	$89.30_{\pm 0.19}$
	Chameleon	$96.61_{\pm 0.30}$	$97.22_{\pm 0.26}$	$99.25_{\pm 0.11}$	$99.72_{\pm 0.06}$	$89.05_{\pm 0.24}$	$96.02_{\pm 0.39}$	$95.90_{\pm 0.18}$	$98.32_{\pm 0.01}$	$82.35_{\pm 0.27}$	$86.99_{\pm 0.57}$
	MOSAIC	$97.98_{\pm 0.05}$	$98.59_{\pm 0.12}$	$99.60_{\pm 0.03}$	$99.76_{\pm 0.01}$	$89.03_{\pm 0.24}$	$97.49_{\pm 0.11}$	$97.02_{\pm 0.25}$	$98.27_{\pm 0.03}$	$83.61_{\pm 0.31}$	$89.98_{\pm 0.13}$
2400	Random	$97.56_{\pm 0.11}$	$98.23_{\pm 0.12}$	$99.56_{\pm 0.04}$	$99.74_{\pm 0.00}$	$89.57_{\pm 0.08}$	$96.97_{\pm 0.09}$	$96.95_{\pm 0.07}$	$98.30_{\pm 0.01}$	$83.95_{\pm 0.31}$	$89.42_{\pm 0.03}$
	Uncertainty	$97.62_{\pm 0.21}$	$98.10_{\pm 0.17}$	$99.36_{\pm 0.10}$	$99.78_{\pm 0.02}$	$89.19_{\pm 0.15}$	$97.07_{\pm 0.22}$	$96.65_{\pm 0.27}$	$98.29_{\pm 0.05}$	$82.53_{\pm 0.48}$	$88.95_{\pm 0.15}$
	Coreset	$97.59_{\pm 0.02}$	$98.53_{\pm 0.06}$	$99.57_{\pm 0.04}$	$99.67_{\pm 0.04}$	$89.79_{\pm 0.24}$	$97.17_{\pm 0.13}$	$97.17_{\pm 0.10}$	$98.31_{\pm 0.02}$	$83.77_{\pm 0.54}$	$89.75_{\pm 0.02}$
	Chameleon	$96.93_{\pm 0.30}$	$97.50_{\pm 0.25}$	$99.37_{\pm 0.05}$	$99.75_{\pm 0.03}$	$88.97_{\pm 0.44}$	$96.37_{\pm 0.24}$	$96.22_{\pm 0.18}$	$98.31_{\pm 0.00}$	$82.39_{\pm 0.45}$	$87.62_{\pm 0.28}$
	MOSAIC	$98.03_{\pm 0.28}$	$98.78_{\pm 0.15}$	$99.62_{\pm 0.02}$	$99.79_{\pm 0.07}$	$89.26_{\pm 0.45}$	$97.59_{\pm 0.27}$	$96.97_{\pm 0.12}$	$98.33_{\pm 0.03}$	$84.02_{\pm 0.10}$	$90.37_{\pm 0.20}$