A. Appendix

A.1. Dataset details

CARLA refuses to spawn agents that collide with the environment, including the ground. To ensure agents are grounded, for any asset that causes a spawn collision, we increase its Z coordinate and try to spawn again . This approach allows us to place every agent on the map, albeit some of the conflicting agents have to 'drop' from above, and consequently we wait for 50 timesteps so those agents can settle. In that duration, the autopilot policy guides the agents to satisfactory positions. After those 50 steps, we then record for another 150 steps and save every 15th frame. The resulting episodes each have ten frames following an initial distribution influenced by Nuscenes and CARLA, and a traffic policy influenced by only CARLA determining the final settled distribution.

We then need the 2D ground truth boxes for each asset. We found the existing suggested approach lacking because it frequently has trouble with occlusions and other challenging scenarios. See below for heuristics we developed to help filter the ground truth boxes. While they are not airtight, the resulting ground truths were qualitatively perceived as more reliable.

- Filter Height: We require that the final $2d$ box is at least 30 pixels. This is in between the easy (40) and medium/hard (25) settings on KITTI [15].
- Max Distance: We require that the ground truth detection not be more than 250 meters away. We enforce this through the use of a depth camera attached to the ego agent.
- Visible Pixel Percent (VPP) and Min Visible Count (MVC): The 2D box is attained by pairing the 3D box with the camera's calibration. With the latter, we get the closest point P to the ego agent. We then get the depth camera's output at the 2D box. VPP asks what percent t of that box is closer than P and filters it if $t \geq 80$, ensuring that at least 20% of the object is not occluded. MVC asks how many pixels q are further than P and filters it if $q < 1300$, ensuring that the occluded object is big enough.

A.2. Supporting charts

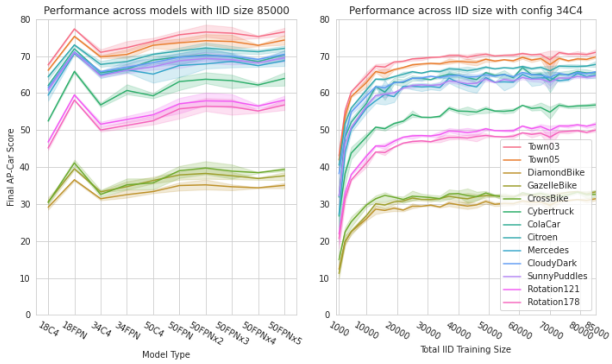

Figure 7: Charts showing increasing both data and model capacity at the same time. The left side ranges over model capacity with maximum IID data size (85000), while the right side ranges over IID data size with a bigger model - 34C4.

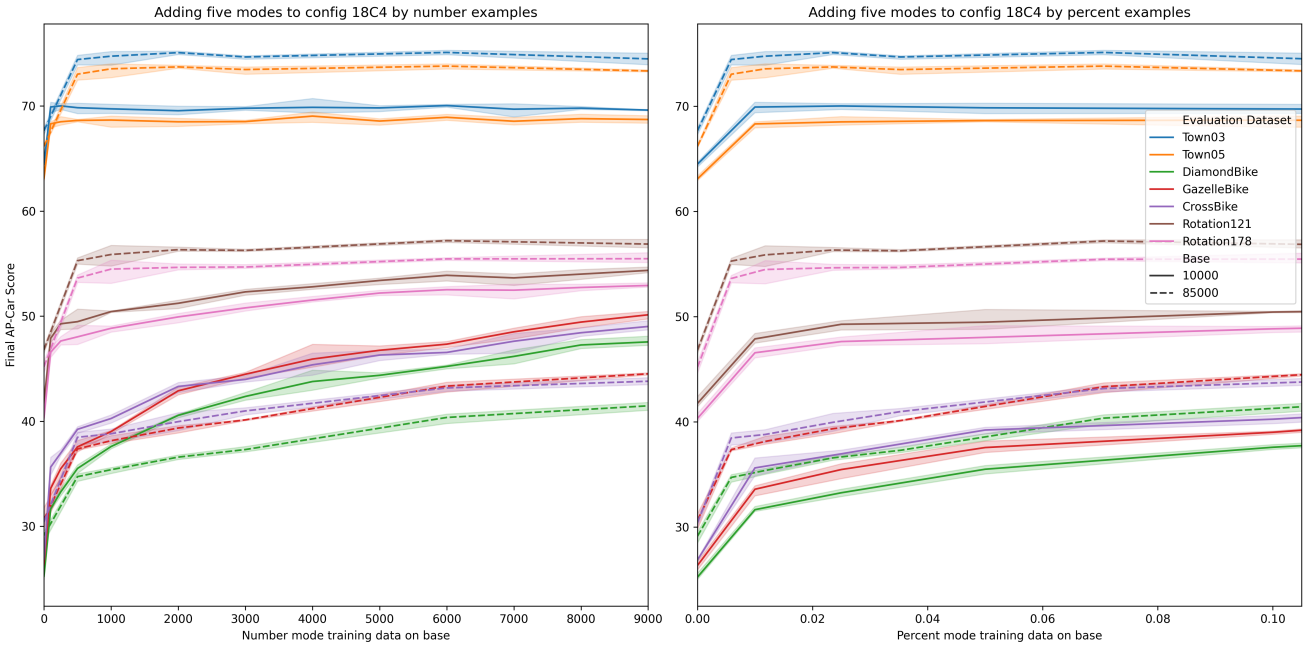

Figure 8: Performance of 18C4 on select test sets when adding mode data from the three bikes, the ColaCar, and the Cybertruck on top of either 10000 or 85000 base IID data. Towards improving the results, these two charts show that it is not the absolute count of the mode data that is important but rather the percent of it relative to the IID data. We see that in how the trendlines for the two bases are only consistent in the percent chart. The other modes are not shown for clarity but it holds in general.

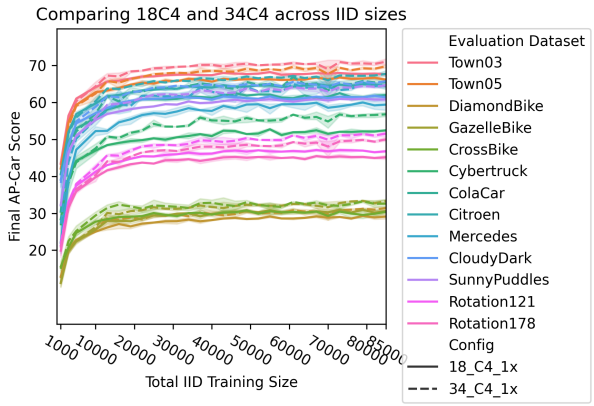

Figure 9: We can see that the model size does matter in that for every group the 34C4 model improves over the 18C4 model. However, the increase is quite small and the data quality and quantity appear to matter much more.

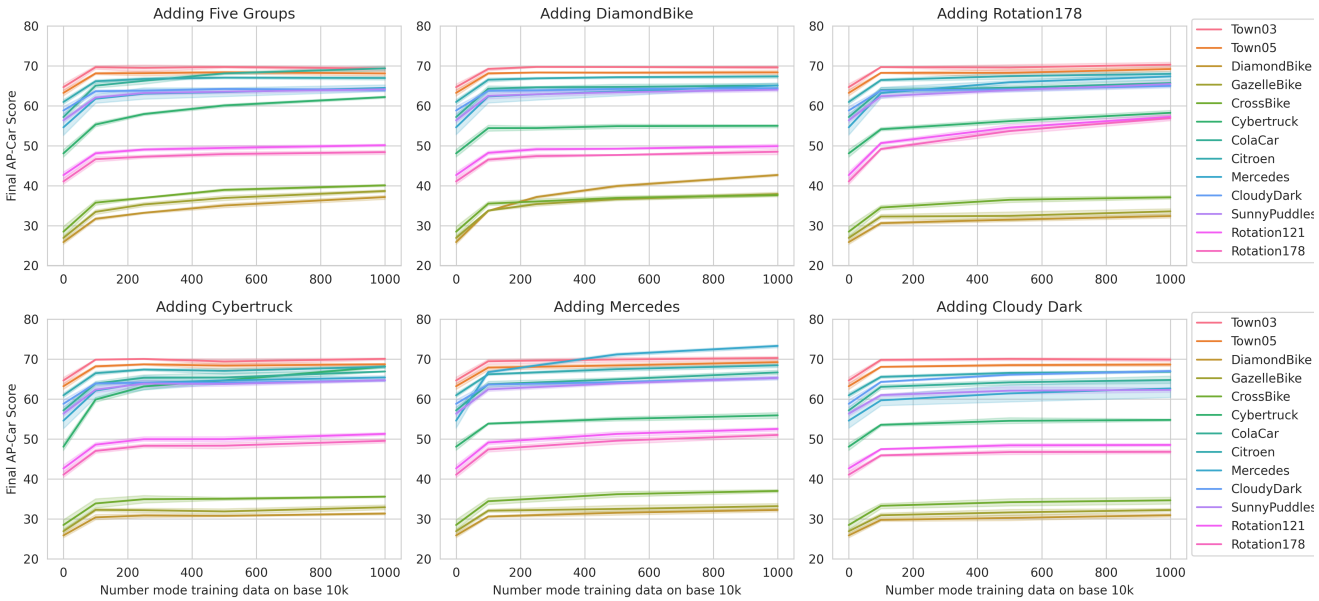

Figure 10: Results adding mode data to the base IID 10000 training set. This is the same as Figure 5 but zoomed into just $[0, 1000]$. The five modes in the top left are the Cybertruck, Cola Car, Diamondback, Gazelle, and Crossbike, each added in equal proportion.

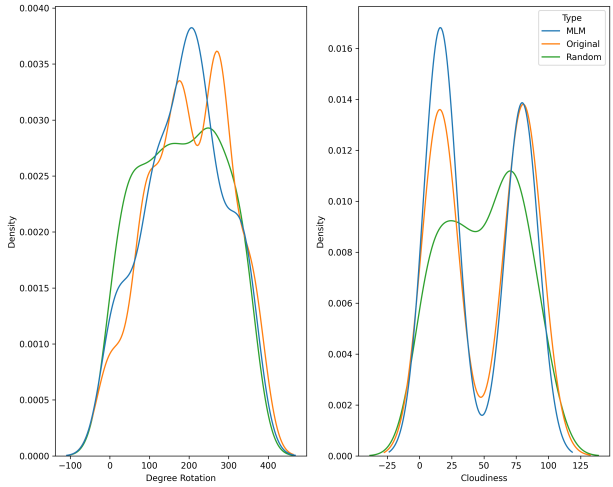

Figure 11: Comparing rotation and weather results for MLM and Random intervention strategies. We see that MLM fits with Original much better than Random does. Further, Random has a much wider berth of possible problematic modes, which is a concern given practical limits to model capacity and data budgets.