Improving Adaptive Cruise Control via Curriculum and Anti-Curriculum Reinforcement Learning

Joshua Glaspey

***Abstract*—This project investigates the application of curriculum and anti-curriculum learning strategies in training a reinforcement learning (RL) based Adaptive Cruise Control (ACC) system. The ACC agent is tasked with maintaining safe following distances behind a lead vehicle while responding to dynamic speed variations of increasing difficulty. I designed a series of lead vehicle speed profiles-categorized as easy, medium, and difficult-using sinusoidal functions with progressively higher amplitude, frequency, and noise to simulate realistic traffic conditions. A Soft Actor-Critic (SAC) agent is trained under three regimes: curriculum learning (easy-to-hard progression), anti-curriculum learning (hard-to-easy regression), and a baseline strategy trained solely on the difficult profile as a control. Experimental results reveal that both curriculum and anti-curriculum learning achieve 100% safe zone adherence on easy and medium profiles, while the baseline struggles on the difficult profile with only 80.2% safety compliance. Curriculum learning in particular achieves smoother control at the cost of higher jerk variance, while anti-curriculum training offers more balanced performance across difficulty levels. These results suggest that progressive difficulty scheduling leads to safer behaviors than naive training on the hardest task alone.**

Link to GitHub Page: https://github.com/jkglaspey/EEL6938-Final-Project-Curriculum-Learning-ACC

I. Introduction

ACC systems are a critical component of modern autonomous driving, enabling vehicles to dynamically adjust their speed to maintain a safe following distance from a lead vehicle in real time. In RL-based implementations of ACC, the agent must simultaneously: minimize speed error relative to the lead vehicle, preserve a safe following distance, and ensure ride comfort by avoiding abrupt accelerations or decelerations. While standard approaches often train agents on a fixed environment or a single difficulty setting, they overlook how the order and complexity of training experiences might affect the final policy's safety, generalization, or robustness. This project builds on the previous framework established in *RL Assignment 2* from the course *Artificial Intelligence for Autonomous Systems* at the *University of Central Florida*, where a similar SAC-based ACC model was trained against a single noisy lead vehicle profile without structured progression.

This project incorporates curriculum learning and anti-curriculum learning—two techniques used in machine learning to structure the order in which training data is presented. In curriculum learning, training begins with simple tasks and gradually advances to more difficult ones. This mimics how humans learn and is designed to build foundational knowledge before introducing complexity. In contrast, anti-curriculum learning inverts this sequence, beginning with the hardest task and regressing toward easier ones. Although counterintuitive, anti-curriculum learning may improve generalization and robustness by forcing the agent to cope with high-variance environments early in training. This addresses any potential concerns of overfitting that curriculum-based learning presents.

To apply these strategies, I designed three synthetic lead vehicle profiles using noisy sine waves with increasing amplitude, frequency, and Gaussian noise. These profiles represent Easy, Medium, and Difficult levels of driving complexity. The RL agent was then trained under three distinct schedules: a curriculum learning schedule (easy $\rightarrow$ medium $\rightarrow$ difficult), an anti-curriculum schedule (difficult $\rightarrow$ medium $\rightarrow$ easy), and a baseline condition on which the agent trained exclusively on the difficult profile. The SAC algorithm was used across all conditions to stay consistent with the prior framework.

The goal of this study is to evaluate whether structured difficulty scheduling during training leads to safer, more responsive, and more comfortable driving policies. Each strategy is assessed using both quantitative performance metrics, such as average reward, speed tracking error, safety zone compliance, and jerk variance-and qualitative visualizations of speed, distance, and comfort over time. This exploration contributes to a deeper understanding of how training sequence influences policy effectiveness in real-world control systems.

II. Methodology

A. Overview

This experiment extends the RL framework established in *Reinforcement Learning Exercise 2* by evaluating the impact of structured training schedules—specifically curriculum and anti-curriculum learning-on the performance of an ACC system. The original system used a single lead vehicle with noise-added sinusoidal behavior derived from a reference speed profile. In this work, that reference speed profile was removed entirely. Instead, the ego vehicle must respond solely to the observed behavior of a dynamic lead vehicle, which follows one of three difficulty-based speed patterns: Easy, Medium, and Difficult.

Each speed profile is defined using a unique sinusoidal function with Gaussian noise to simulate realistic traffic behavior. These profiles vary in frequency and amplitude to introduce progressively greater complexity. The Easy profile exhibits slow, shallow oscillations with minimal noise; the Difficult profile uses higher-frequency sine waves with stronger stochastic disturbances, making smooth tracking more challenging. The equations for each profile are shown in Section 2.2.

The experiment tests three training strategies:

- **Curriculum Learning** - The model trains sequentially from Easy to Medium to Difficult profiles.
- **Anti-Curriculum Learning** - The model trains from Difficult to Medium to Easy.
- **Baseline** - The model trains solely on the Difficult profile.

Each training schedule uses 100,000 timesteps, divided evenly across the difficulty segments. The models are evaluated using a combined 3,600-step test set that concatenates the Easy, Medium, and Difficult profiles in order. The objective is to determine whether the order in which complexity is introduced during training influences the agent's ability to learn safe, smooth, and accurate vehicle-following behavior.

B. Environment Design

The simulation environment models a simplified one-dimensional driving scenario in which an ego vehicle is trained to follow a lead vehicle while maintaining a safe following distance. The simulation operates with discrete-time dynamics at a fixed timestep of $\Delta t = 1.0$ second. At each step, the environment returns a three-dimensional observation vector containing:

- the ego vehicle's current speed v_{ego},
- the lead vehicle's current speed v_{lead},
- and the relative distance between them $d = x_{lead} - x_{ego}$.

The ego vehicle applies a continuous acceleration value within safe physical bounds: the applied acceleration is clamped between $-2.0 m/s^2$ and $+2.0 m/s^2$. Velocity is non-negative and is clamped at zero if the applied acceleration would result in reversal. At the start of each episode, the ego vehicle is initialized at a fixed displacement of -17.5 meters relative to the lead and begins at the same speed as the lead vehicle's initial velocity in the selected episode. The initial alignment in speed prevents reward spikes due to immediate mismatch, and the displacement allows the vehicle to enter the safe zone naturally.

Training data is generated from three difficulty levels of lead vehicle speed profiles: Easy, Medium, and Difficult. Each profile consists of a 1200-timestep sequence of speed values constructed as noisy sine waves:

- **Easy:** $v(t) = 10 + 0.2sin(0.01t) + \mathcal{N}(0, 0.1^2)$
- **Medium:** $v(t) = 10 + sin(0.05t) + \mathcal{N}(0, 0.5^2)$
- **Difficult:** $v(t) = 10 + 3sin(0.1t) + \mathcal{N}(0, 1^2)$

Each lead speed profile is stored in a separate CSV file and transformed into a position profile using a discrete cumulative sum:

$$x_{lead}(t) = \sum_{i=0}^{t} v_{lead}(i)$$

This ensures consistent forward motion and allows the relative distance $d(t)$ to be calculated at each step. For training, each profile is split into 12 non-overlapping 100-step segments using a chunking function. At the beginning of each episode, one chunk is randomly selected based on the current difficulty schedule and assigned as the lead vehicle trajectory.

The environment enforces a safe following distance of 5 to 30 meters. If the ego vehicle remains within this zone, no penalty is applied. Deviations below 5 meters incur a sharply increasing penalty, while distances beyond 30 meters are penalized linearly. These penalties are encoded in the reward function Section 2.5, which also includes penalties for jerk and acceleration magnitude to encourage comfort and mechanical feasibility.

C. Learning Algorithm

All experiments were conducted using the SAC algorithm, a deep RL method well-suited for continuous control. The default configuration of hyperparameters was used, which is provided below.

- Learning Rate: 3e-4
- Batch Size: 256
- Buffer Size: 200,000
- Tau: 0.005
- Gamma: 0.99
- Entropy Coefficient: 'auto'

For the episodes, a chunk size of 100 was used to create 12 episodes per speed profile.

D. Difficulty Scheduling

To evaluate how the order of training difficulty influences performance, the agent was trained using three distinct scheduling strategies: curriculum learning, anti-curriculum learning, and a baseline. Each strategy determines how the lead vehicle's difficulty level changes over the course of training.

In the **curriculum** learning schedule, the agent begins training on the easiest profile and progresses to harder ones. Specifically, training is divided into three equal segments of 33,333 steps:

- Steps 0-33,333: Easy profile
- Steps 33,334-66,665: Medium profile
- Steps 66,666-99,999: Difficult profile

In contrast, the **anti-curriculum** learning schedule reverses this order, beginning with the most challenging environment and regressing to easier tasks:

- Steps 0-33,333: Difficult profile
- Steps 33,334-66,665: Medium profile
- Steps 66,666-99,999: Easy profile

The **baseline** strategy omits scheduling entirely and trains on the Difficult profile for all 100,000 steps.

All schedules use the same chunked episode format, where the lead vehicle profile is segmented into 100-step sub-trajectories that are randomly sampled during training. This episodic segmentation provides a diverse set of sub-scenarios while maintaining consistent difficulty within each schedule phase.

The goal of this comparison is to determine whether introducing complexity gradually (as in curriculum learning) or exposing the agent to high difficulty early (as in anti-curriculum learning) results in better overall performance, safety, and generalization under varied conditions.

E. Reward Function

The agent's objective is to maintain a safe following distance, track the lead vehicle's speed, and minimize abrupt changes in motion. To support this, the reward function penalizes deviation from the lead vehicle's speed, unsafe distances, high jerk, and high acceleration magnitudes. The overall reward at each timestep is defined as the negative sum of these four error components:

$$\text{Reward} = -(|v_{ego} - v_{lead}| + 3 * DistanceError(d) + 0.1 * |jerk_{ego}| + 0.05 * |acceleration_{ego}|)$$

Here, v_{ego} and v_{lead} represent the speeds of the ego and lead vehicles, respectively, and d is the current distance between them. The distance error is a piecewise function defined as:

$$\text{DistanceError}(d) = \begin{cases} 100, & \text{if } d \leq 0 \\ 3 \cdot |5 - d|, & \text{if } 0 < d < 5 \\ |30 - d|, & \text{if } d > 30 \\ 0, & \text{otherwise} \end{cases}$$

This formulation imposes a heavy penalty for collisions ($d \leq 0$), a steep penalty for following too closely ($0 < d < 5$), and a softer penalty for falling behind the lead vehicle too far ($d > 30$). When the ego vehicle remains within the designated safe following zone (5-30 meters), no penalty is applied for distance.

The jerk is computed as the first-order difference of consecutive acceleration values, which encourages the agent to make smooth changes in velocity. The inclusion of an explicit acceleration penalty prevents the policy from issuing overly aggressive actions, which promotes comfort.

This reward function allows the agent to learn policies that balance responsiveness, safety, and ride comfort while generalizing across dynamic lead vehicle behaviors of varying difficulty.

F. Evaluation Metrics

To evaluate the ACC system, I track ten key metrics during testing. For these definitions, let r_i represent the reward at time step i, $v_{\text{ego},i}$ represent the ego vehicle's speed, $v_{\text{lead},i}$ represent the lead vehicle's speed, d_i be the distance between the vehicles, a_i the acceleration of the ego vehicle, j_i the jerk, and N the total number of testing steps.

1) **Average Reward:** The average cumulative reward received by the ego vehicle during testing. Higher values indicate better policy performance.

$$\text{Average Reward} = \frac{1}{N}\sum_{i=1}^{N} r_i$$

2) **Mean Absolute Error (MAE):** The mean absolute deviation between the ego and lead vehicle speeds. Lower values indicate more accurate speed tracking.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |v_{\text{ego},i} - v_{\text{lead},i}|$$

3) **Root Mean Squared Error (RMSE):** The square root of the mean squared speed tracking error. More heavily penalizes large deviations.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (v_{\text{ego},i} - v_{\text{lead},i})^2}$$

4) **R^2 Score:** The proportion of variance in the lead vehicle's speed that is captured by the ego vehicle's speed.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(v_{\text{ego},i} - v_{\text{lead},i})^2}{\sum_{i=1}^{N}(v_{\text{lead},i} - \bar{v}_{\text{lead}})^2}$$

5) **Mean Distance:** The average distance maintained by the ego vehicle from the lead vehicle.

$$\text{Mean Distance} = \frac{1}{N}\sum_{i=1}^{N} d_i$$

6) **Safe Zone Percentage:** The percentage of time steps where the ego vehicle's distance from the lead vehicle was within the safe zone of 5–30 meters.

$$\text{Safe Zone } \% = \frac{1}{N}\sum_{i=1}^{N} 1_{[5 \leq d_i \leq 30]} \times 100$$

7) **Minimum Distance:** The smallest distance observed between the ego and lead vehicle during testing.
8) **Maximum Distance:** The largest distance observed between the ego and lead vehicle during testing.
9) **Mean Jerk:** The average jerk magnitude experienced by the ego vehicle, indicating ride smoothness.

$$\text{Mean Jerk} = \frac{1}{N}\sum_{i=1}^{N} |j_i|$$

10) **Jerk Variance:** The variance in jerk values, measuring the consistency of the vehicle's comfort profile.

$$\text{Jerk Variance} = \frac{1}{N}\sum_{i=1}^{N} (j_i - \bar{j})^2$$

III. Results

The trained agents were evaluated on a fixed 3,600-step test sequence composed by concatenating the Easy, Medium, and Difficult lead vehicle profiles. Each training strategy—Curriculum, Anti-Curriculum, and Baseline—was run independently, and the resulting policies were evaluated using the metrics defined in Section 2.6. The results below reflect the agent's ability to generalize across varying difficulty levels.

For reference, Figure 1 shows the Easy, Medium, and Difficult speed profiles as their respective sine waves. These were consistent between experiments and reflect the variations in speed and comfort between simple and complex driving.

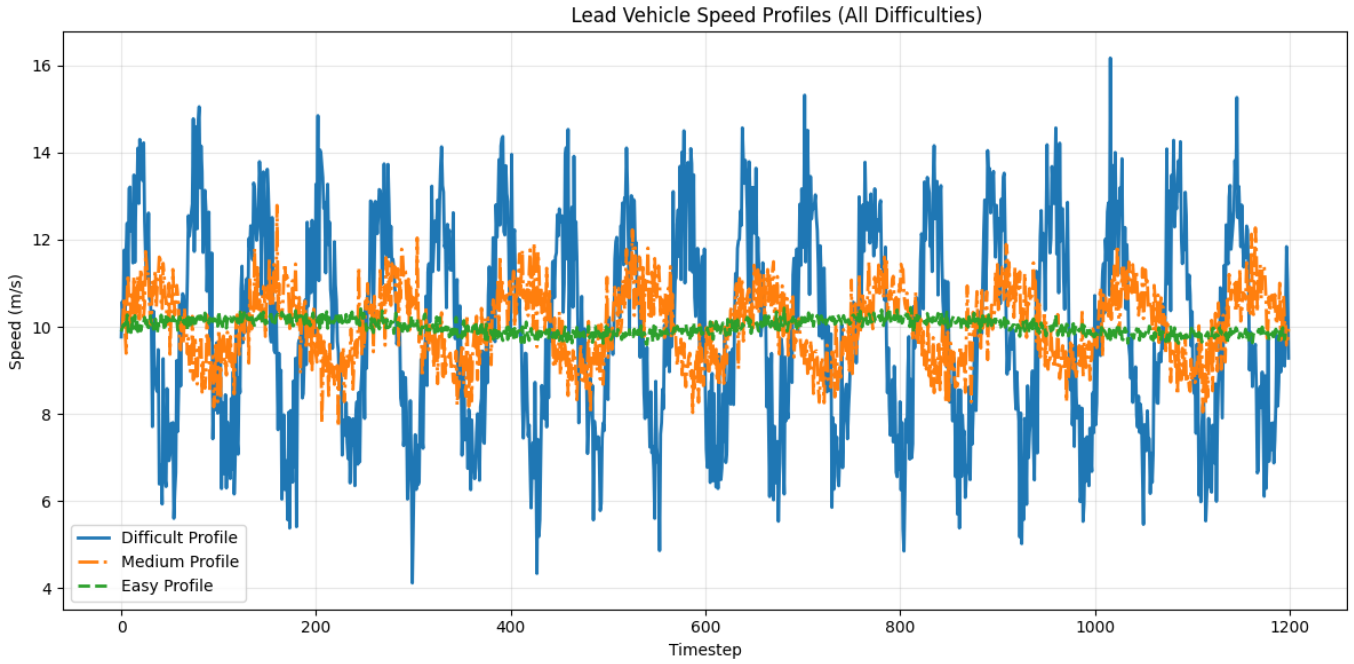

Fig. 1: Lead Profile Plot (Easy, Medium, Difficult). Visualizes the speeds of each individual training strategy and emphasizes the differences in complexity.

A. *Quantitative Results*

The comprehensive set of metrics were calculated and reported across four scenarios: the Easy profile, Medium profile, Difficult profile, and an Overall profile that concatenates all three difficulty levels. Each of these profiles was tested under three training regimes: curriculum learning, anti-curriculum learning, and a baseline trained solely on the Difficult profile. The results are summarized in Tables 1 through 4.

Table I reports the results for the Easy profile section of the test profile, highlighting model behavior under stable lead vehicle oscillations. Table II presents metrics for the Medium profile, which introduces higher variability in lead speed. Table III summarizes performance under the Difficult profile, which imposes sharp speed fluctuations and increased noise. Finally, Table IV aggregates all test segments into a single evaluation window and provides an overall performance comparison.

To improve readability, each group of evaluation metrics is color-coded based on the behavior category it represents:

- **Blue rows** represent Speed Tracking Metrics, including MAE, RMSE, and the R^2 Score. These metrics reflect how well the ego vehicle mimics the speed of the lead vehicle.
- **Green rows** represent Distance Maintenance Metrics, measuring how well the ego vehicle maintains a safe and desirable distance from the lead vehicle. This includes the mean, min, and max following distances, and the percentage of time spend within the safe zone.
- **Red rows** represent Comfort Metrics, based on jerk calculations. These provide insight into the smoothness of the ride experienced by passengers in the ego vehicle, measured through the mean, maximum, and variance of jerk.

Easy Profile	Curriculum	Anti-Curriculum	Baseline
Avg. Reward	-0.147	-0.126	-0.103
MAE	0.129 m/s	0.117 m/s	0.107 m/s
RMSE	0.364 m/s	0.370 m/s	0.316 m/s
R2 Score	-0.174	-0.181	0.113
Mean Distance	13.942 m	8.647 m	23.868 m
Safe Zone %	100.00%	100.00%	100.00%
Min Distance	6.356 m	5.589 m	8.036 m
Max Distance	14.590 m	9.198 m	24.738 m
Mean Jerk	-0.000 m/s^3	-0.000 m/s^3	-0.000 m/s^3
Max Jerk	2.379 m/s^3	1.840 m/s^3	1.244 m/s^3
Jerk Variance	0.076 m^2/s^6	0.040 m^2/s^6	0.003 m^2/s^6

TABLE I: Performance of all training strategies during the first 1200-timesteps of the testing speed profile (easy speed profile). The underlined results are the best score of their respective rows.

B. *Visualizations*

To supplement the quantitative evaluation, five visualizations were generated for each of the three training strategies. These plots provide a qualitative view of the ego vehicle's performance over the full 3600-timestep evaluation sequence. Each figure corresponds to a specific performance dimension-distance maintenance, speed tracking, control smoothness, or learning behavior-and visualizes the response of each policy to the sequence of Easy, Medium, and Difficult lead vehicle speed profiles.

Medium Profile	Curriculum	Anti-Curriculum	Baseline
Avg. Reward	-0.601	-0.615	-0.457
MAE	0.501 m/s	0.532 m/s	0.437 m/s
RMSE	0.630 m/s	0.662 m/s	0.548 m/s
R2 Score	0.459	0.424	0.582
Mean Distance	14.141 m	8.855 m	24.196 m
Safe Zone %	100.00%	100.00%	100.00%
Min Distance	11.069 m	6.764 m	20.474 m
Max Distance	16.757 m	11.289 m	28.149 m
Mean Jerk	0.000 m/s^3	0.000 m/s^3	-0.000 m/s^3
Max Jerk	3.001 m/s^3	2.557 m/s^3	0.922 m/s^3
Jerk Variance	0.914 m^2/s^6	0.605 m^2/s^6	0.041 m^2/s^6

TABLE II: Performance of all training strategies during the second 1200-timesteps of the testing speed profile (medium speed profile). The underlined results are the best score of their respective rows.

Difficult Profile	Curriculum	Anti-Curriculum	Baseline
Avg. Reward	-1.247	-1.228	-1.929
MAE	1.060 m/s	0.988 m/s	0.977 m/s
RMSE	1.330 m/s	1.242 m/s	1.221 m/s
R2 Score	0.675	0.709	0.728
Mean Distance	14.230 m	9.022 m	24.413 m
Safe Zone %	100.00%	98.10%	80.20%
Min Distance	6.838 m	3.271 m	14.294 m
Max Distance	20.994 m	14.846 m	34.941 m
Mean Jerk	0.000 m/s^3	0.000 m/s^3	0.000 m/s^3
Max Jerk	3.888 m/s^3	3.834 m/s^3	2.125 m/s^3
Jerk Variance	2.977 m^2/s^6	1.810 m^2/s^6	0.358 m^2/s^6

TABLE III: Performance of all training strategies during the last 1200-timesteps of the testing speed profile (difficult speed profile). The underlined results are the best score of their respective rows.

The **following distance** plots reveal whether the ego vehicle maintained a safe and stable separation from the lead vehicle. These plots complement the metrics like Mean Distance, Min/Max Distance, and Safe Zone %, providing a temporal context for how and when deviations occurred. The **speed comparison** plots directly visualize how closely the ego tracked the lead vehicle's velocity, supporting the MAE, RMSE, and R^2 metrics. The **speed difference** plots further isolate these deviations to emphasize tracking error magnitude over time.

The **reward penalty over time** plots visualize the stability and learning progression of each training strategy. While the numerical Average Reward captures the final value, these plots reveal periods of instability or stagnation that are otherwise obscured. Finally, the **jerk analysis** plots offer a time-resolved view of ride comfort, complementing the Max Jerk and Jerk Variance metrics reported in the tables. Together, these figures provide a holistic understanding of system behavior that cannot be conveyed through aggregate metrics alone.

Overall Profile	Curriculum	Anti-Curriculum	Baseline
Avg. Reward	-0.665	-0.656	-0.830
MAE	0.564 m/s	0.546 m/s	0.507 m/s
RMSE	0.875 m/s	0.840 m/s	0.794 m/s
R2 Score	0.635	0.657	0.700
Mean Distance	14.105 m	8.841 m	24.159 m
Safe Zone %	100.00%	99.40%	93.40%
Min Distance	6.356 m	3.271 m	8.036 m
Max Distance	20.994 m	14.846 m	34.941 m
Mean Jerk	0.000 m/s^3	0.000 m/s^3	0.000 m/s^3
Max Jerk	3.888 m/s^3	3.834 m/s^3	2.125 m/s^3
Jerk Variance	1.322 m^2/s^6	0.818 m^2/s^6	0.134 m^2/s^6

TABLE IV: Performance of all training strategies during all 3600-timesteps of the testing speed profile. The underlined results are the best score of their respective rows.

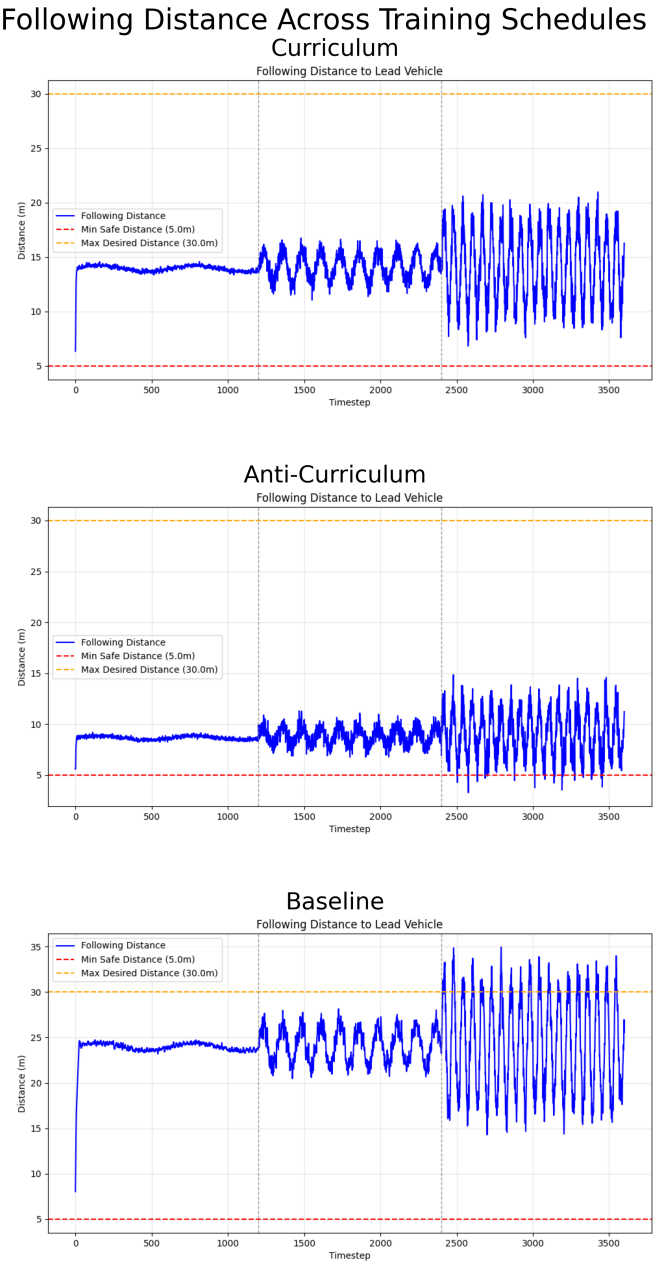

Fig. 2: Following distance between the ego and lead vehicle for each training schedule across the 3600-timestep test. The dotted red and orange lines represent the minimum (5m) and maximum (30m) desired distances, respectively.

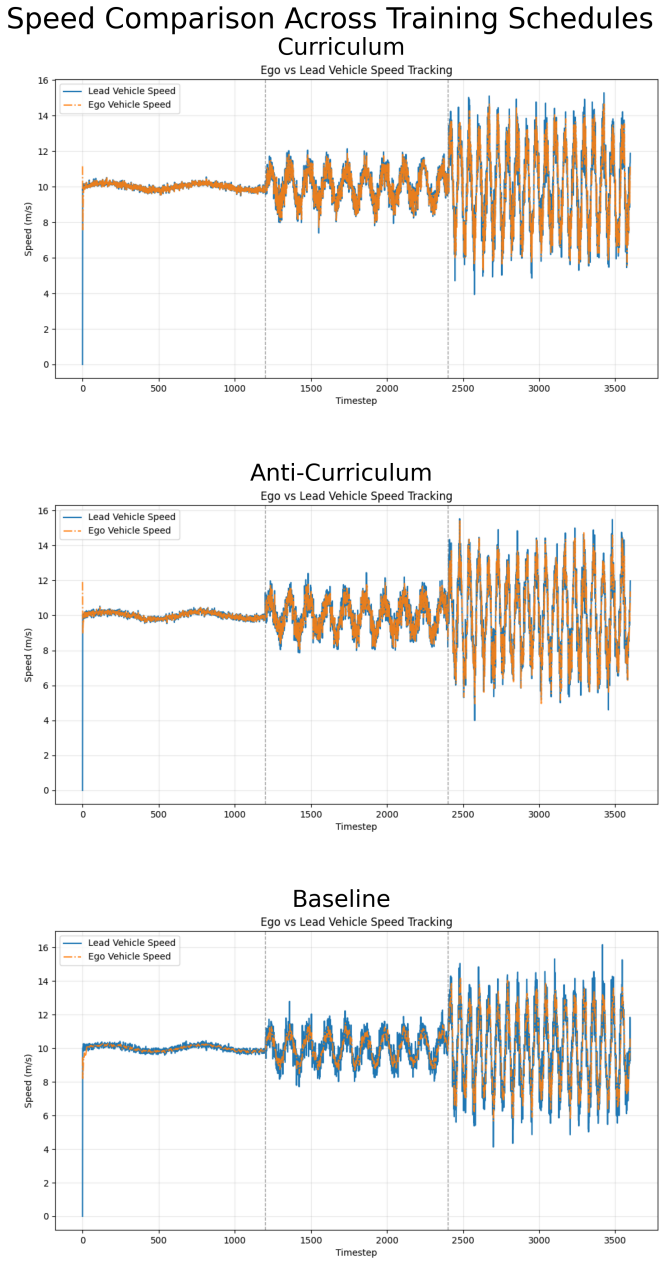

Fig. 3: Overlay of ego and lead vehicle speeds across all testing timesteps, shown for each training strategy. The orange line is the ego vehicle, and the blue line is the lead vehicle. Close tracking reflects lower MAE and RMSE errors.

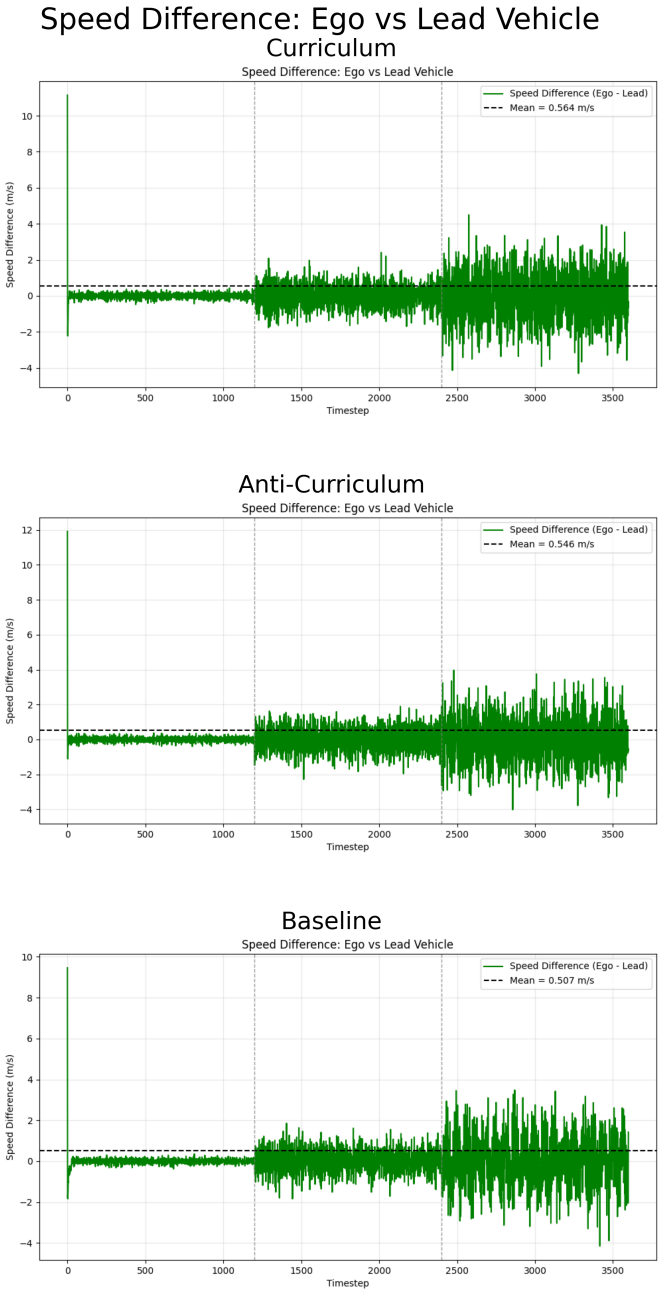

Fig. 4: Instantaneous difference between ego and lead vehicle speeds throughout the test. Smaller absolute values indicate better velocity tracking.

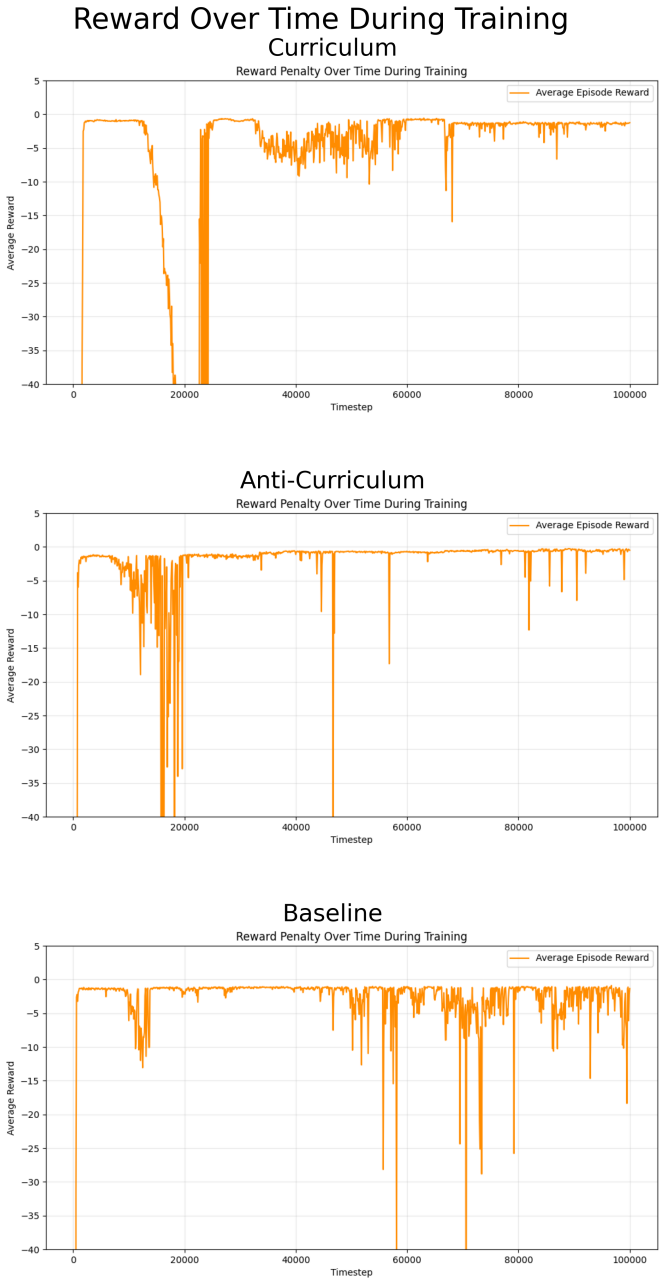

Fig. 5: Rolling average of episode rewards during training for each strategy. Periods of instability or convergence plateaus are observable from the dips and recoveries. The y-axis was clamped at -40 to emphasize small reward oscillations.

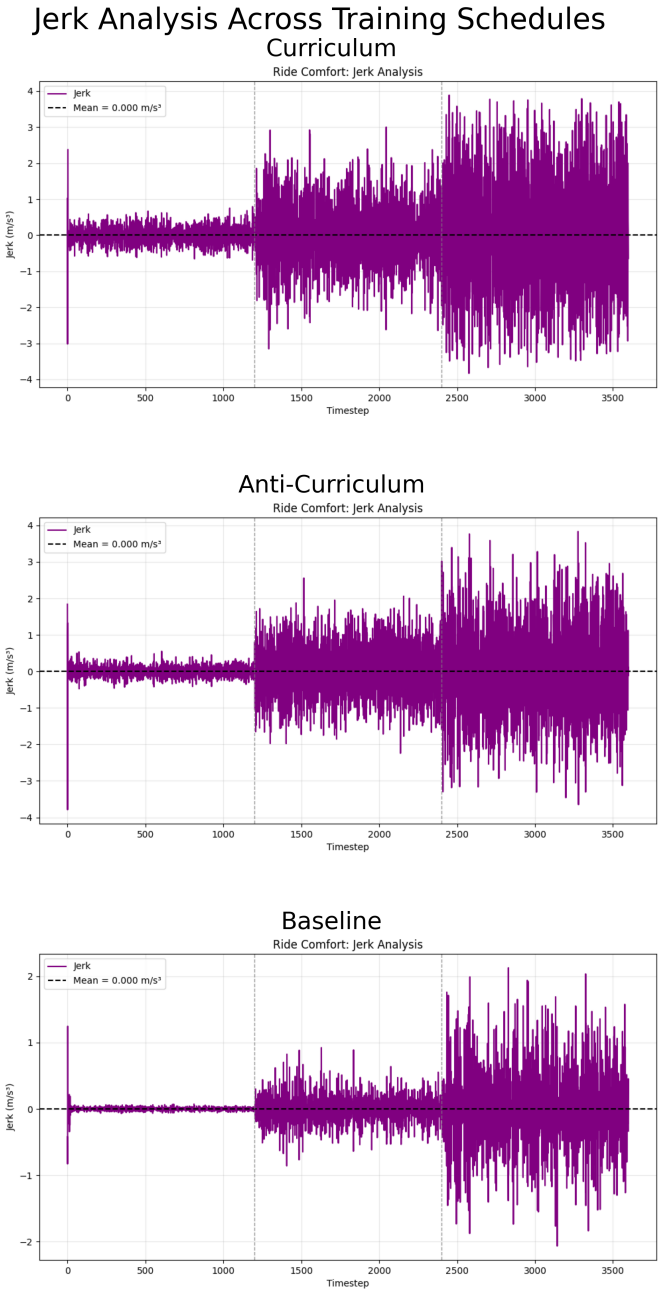

Fig. 6: Magnitude of jerk (rate of acceleration change) throughout the test episode. Lower magnitudes and lower variance suggest improved ride comfort.

C. Observations

The performance metrics and corresponding visualizations provide detailed insight into how each training strategy responds to the Easy, Medium, and Difficult lead vehicle profiles. This section examines each metric category in depth—Speed Tracking, Distance Maintenance, and Ride Comfort—to highlight patterns and trade-offs between curriculum, anti-curriculum, and baseline training regimes.

1) Speed Tracking Metrics (MAE, RMSE, R^2 Score): The baseline model consistently achieved the best MAE and RMSE

values across all difficulty segments, indicating the smallest speed deviations from the lead vehicle. These results are further validated by the speed comparison plots (Figure 3), where the ego vehicle's velocity almost perfectly overlays the lead vehicle's profile under the baseline. Despite this precision, the R^2 Score results are less straightforward. While the baseline exhibits the best R^2 values for the Easy and Medium profiles, it narrowly outperforms the other models in the Difficult segment as well. However, the curriculum model maintains better tracking consistency across difficulties, evidenced by its relatively stable MAE and RMSE. Anti-curriculum learning performs moderately in all three tests, with marginally worse speed tracking than the baseline. These trends are corroborated by the speed difference plots (Figure 4), where the baseline's curve shows the lowest mean deviation but also larger outliers under Difficult conditions, suggesting riskier, abrupt corrections.

2) Distance Maintenance Metrics (Mean/Min/Max Distance, Safe Zone %): Curriculum learning shows the most balanced distance maintenance behavior. It maintains a safe yet moderate following distance across all segments and achieves 100% Safe Zone adherence in all test conditions. Anti-curriculum training, while also achieving full compliance in the Easy and Medium profiles, exhibits reduced safety in the Difficult section, dipping to 98.1% due to occasional close-following behavior, as confirmed by its minimum distance of 3.271 m. The baseline model performs poorly in this category, especially under the Difficult profile, where the Safe Zone adherence drops to 80.2%. The following distance plots (Figure 2) reveal that the baseline model tends to hover near the upper safety boundary (30 m), leading to inflated Mean and Max Distance values. In contrast, the curriculum model maintains a more centralized distance band, reflecting consistent control.

3) Ride Comfort Metrics (Mean Jerk, Max Jerk, Jerk Variance): Curriculum learning results in the highest jerk magnitudes and variability, especially in the Difficult segment where Max Jerk exceeds 3.8 m/s^3 and Jerk Variance reaches nearly 3.0 m^2/s^6. This trend is visible in the jerk analysis plots (Figure 6), which show large oscillations in the final third of the test. This is expected, as the curriculum-trained agent concludes training with the most volatile lead profile. Anti-curriculum training exhibits lower jerk values and smoother motion transitions, especially toward the end of testing, which aligns with its training trajectory concluding with the Easy profile. The baseline model, while achieving the lowest jerk variance and maximum jerk across all conditions, sacrifices distance maintenance and stability to achieve these comfort metrics. Overall, there is a clear trade-off between tracking precision, safety, and comfort, with curriculum learning emphasizing safety, baseline prioritizing tracking accuracy, and anti-curriculum providing a middle ground.

4) Reward Trends: Average reward values correlate strongly with Safe Zone adherence and tracking precision. Curriculum and anti-curriculum strategies yield similar overall rewards (-0.665 and -0.656, respectively), while the baseline performs worse at -0.830. The reward penalty plots (Figure 5) reveal that curriculum learning has a structured pattern of instability and recovery, consistent with scheduled difficulty transitions. Anti-curriculum learning stabilizes early but exhibits smaller fluctuations throughout, reflecting its adaptation from difficult to easy tasks. The baseline model, by contrast, undergoes more frequent reward dips, suggesting difficulties in adapting exclusively to high-difficulty profiles.

In summary, the experiments reveal that curriculum learning prioritizes safety at the cost of comfort, anti-curriculum learning strikes a moderate balance across all objectives, and baseline training optimizes for speed tracking but suffers in distance safety and adaptability. These outcomes reinforce the importance of difficulty scheduling in shaping the behavioral trade-offs of RL-based control systems.

IV. Conclusions

This work examined the effect of difficulty scheduling strategies-specifically curriculum and anti-curriculum learning-on the performance of a RL-based ACC system. By introducing structured lead vehicle profiles of increasing complexity, the study aimed to determine whether training sequence influences the safety, responsiveness, and comfort of learned driving behaviors.

The results demonstrate that the choice of training schedule has a measurable impact on agent behavior. Curriculum learning led to superior distance maintenance, consistently maintaining safe following gaps across all difficulty levels and achieving 100% safe zone adherence throughout. Anti-curriculum learning, while slightly less stable in high-difficulty conditions, offered the most balanced performance overall, achieving strong speed tracking while maintaining reasonable ride comfort. In contrast, baseline training on the difficult profile yielded the best speed tracking metrics but frequently violated safety constraints, especially under high-noise conditions.

The analysis of visualizations and performance metrics revealed trade-offs inherent in each strategy. Curriculum learning emphasized safety but exhibited higher jerk variance. Anti-curriculum learning produced smoother control transitions and more consistent reward progression. Baseline training offered strong tracking fidelity but suffered from poorer generalization to comfortable and safe driving.

Overall, these findings suggest that structured progression in training difficulty enhances the generalization, safety, and comfort of reinforcement learning-based control policies. Curriculum and anti-curriculum approaches both outperform naive training on difficult tasks in at least one key dimension of performance. These insights reinforce the potential of difficulty scheduling as a design lever for more robust and reliable autonomous systems, especially in domains where environmental complexity can be simulated. Future work may explore adaptive or data-driven scheduling strategies, as well as potential multiple-dimension lead vehicle behavior (such as lane switching).