Reviewer #1: This study proposes a deep reinforcement learning-based car-following model for autonomous vehicles. Overall, this study is straightforward, and the topic is interesting.

 Regarding traffic flow modeling and learning-based models in TR-B, the reviewer would like to see more novel methodology parts (e.g., models, properties, and algorithms). For example, Yuan, et al. (2021) proposed a new macroscopic traffic flow model with physics models and Gaussian process (GP). New theorems are provided to show that the proposed shadow-GP is learnable and the feasibility of employing the physical knowledge regularization. Ahamed, et al. (2021) proposed a DQN-based approach for urban delivery. New heuristics are provided to guide specific actions. The computational complexities of the heuristics and the DQN training are analytically investigated. Zhou, et al. (2019) presented a robust car-following control strategy under uncertainty for connected and automated vehicles (CAVs). To analyze the stability of the proposed control system, they mathematically proved a sufficient and necessary condition for local stability and sufficient conditions for robust string stability.

However, the methodology part in this paper is the common knowledge that has been used in existing studies. The authors revised the reward function and traffic environment of the free-driving policy based on a widely used actor-critic network framework, and then used the existing Deep Deterministic Policy Gradient (DDPG) Algorithm for training. There are no mathematical analyses to investigate the feasibility, stability, and computational complexity of the proposed model and algorithm. Therefore, the reviewer suggests the authors make serious revisions to justify the methodological contributions or submit this manuscript to a more application oriented journal.

Dear reviewer

Thank you for your feedback. The main critique addresses the missing methodology parts in this study. However, from our perspective, we see the following methodological contribution.

We propose the first**complete RL based car-following** **model** characterized by:

**M1. Consideration of all kinds of car-following events**

**State-of-the-art:**To our knowledge, all existing RL-based car-following models do not consider all kinds of car-following events, especially safety-critical events, such as full-braking of the leader. Since in RL, the model learns its behavior in a simulative training environment, the design of the simulation is crucial. All reviewed studies model the leading vehicle's trajectories in a simulation environment based on: (i) real driving data, e.g., NGSIM, (ii) standard driving cycles, e.g., the Standard European Driving Cycle, or (iii) even constant speeds. However, such 'usual' driving data is lacking in situations that are safety-critical, such as full-braking of the leader vehicle, resulting in a training environment that covers just a small part of the distribution of possible car-following states. This leads to car-following models suitable for this particular kind of training data, but they cannot be generalized for different kinds of car-following scenarios that are not reflected in training.

**Our approach:**We aim to develop a car-following model that considers all kinds of car-following events, especially safety-critical states, such as the follower reacting to high decelerations of the leader vehicle. We designed an RL training environment based on stochastic processes, in order to cover the whole distribution of possible car-following states. To test the trained model's out-of-distribution robustness (Träuble et al., 2021), we performed validation in different scenarios that have never been seen in training. This includes experimental car-following scenarios recorded in urban and peripheral arterials in Napoli. Moreover, we bring the trained model to its limits by synthetic scenarios where we manipulate the leader vehicle to accelerate and decelerate extremely, including a sudden full-braking maneuver to a standstill. Since there is no analytical way of showing the robustness of machine learning models in general, the best one can do is demonstrating string stability as well as a plausible and crash free behavior in extreme situations that the model has never seen before.

Summarized, we propose the first RL-based car-following model that considers safety-critical events, including a model validation in extreme situations that are not part of the training procedure.

**M2. Consideration of free-driving, car-following, and the smooth transition between the two**

To our knowledge, there exists no RL-based model reflecting not only car-following situations but also free-driving and a transition between the two. However, since too many objectives at once can lead to performance losses, a known problem in Multi-Objective RL (Liu et al., 2014), we propose a synchronized modular RL architecture that separates the objective of free-driving (keeping to the speed limit) and the objective of car-following (driving safely, efficiently and comfortably) into two policies. However, these policies are not strictly separated, but they run both in parallel, each computing an acceleration value each time step, resulting in a model that distinguishes between free-driving and car-following states implicitly.

In the following, we refer to the other comments to improve this study.

1. The free-driving policy only considers the acceleration part. However, there is another scenario in free-driving traffic that vehicles need to decelerate when the speed limit drops. The authors should also consider a deceleration part in the free-driving policy.

We thank the reviewer for this note and implemented and trained the free-driving policy with an additional reward part that reflects deceleration when driving above the speed limit.

1. In the car-following policy, the authors said "when gt exceeds gmax or there is no leader, g is set to gmax". If there is no leader, it seems the vehicle is in a free-driving scenario. Then, this should not be a car-following policy.

As stated in M2, our synchronized modular RL system does not really distinguish between free-driving and car-following scenarios. Both policies always compute an acceleration value, and just the minimum of both accelerations is the control input for the vehicle. That means the car-following policy also has to output an acceleration value, even when there is no leader vehicle in front. Therefore, we use g\_max as a technical solution to limit the range of g\_t that must be considered in training, and we set its value to a reasonably high value.

1. The authors used several datasets in this study. One dataset is field data, and the others are simulated data. Why did the authors use these datasets? It seems the field dataset is enough. Why did the authors use simulated datasets? How did the authors use these datasets for training and validation?

We hope that we could clarify the choice of different datasets for training and validation in the M1.

1. The authors compared the proposed model with the IDM. But the IDM is used for human-driven vehicles, not for autonomous vehicles. The authors should compare the proposed model with existing RL based models.

The IDM does not consider reaction times and human perception failures. Therefore the IDM can indeed be seen as an ACC model. Furthermore, our RL model uses the same input and output as the IDM. Therefore, from our perspective, the IDM is a suitable model to compare the performance of our RL model with.  
Furthermore, we did not consider other RL-based car-following models in the validation because our model is the only complete RL-based car-following model considering free-driving and the transition to car-following as well as a bigger range of possible decelerations in the conceptualization of the RL system (see M1 and M2). As the existing RL-based approaches do not consider these features, we cannot perform a comparison with these models.  
Our contribution is not another RL-based car-following model being compared with existing RL-based models but rather to propose the first complete RL-based model.

1. The organization is terrible. For example, table 1 and figure 2 show the results of the car-following policy before it is mentioned.

To fix these issues, we restructured the paper.

1. In the last paragraph on page 2, the authors mentioned some RL based AV car-following models. However, one study (Zhou, et al., 2017) is a supervised learning-based model for human-driven vehicle trajectory prediction. It should not be considered here. The authors should carefully review these studies.

This section of the introduction is not about RL-based car-following models but about data-driven approaches using Deep Learning methods. These methods include supervised learning in general, so from our perspective, the study of Zhou et al. (2017) does fit in this section of the introduction. We restructured the introduction in order to stress the difference between studies based on Deep Learning methods in general and the ones based on Deep RL in particular.

1. There is a wrong formulation in equation (8). The reviewer believes that "d" is supposed to be "Δt". Please carefully proofread this paper.

We thank the reviewer for spotting this typo and have carefully proofread the revised paper.

1. In section 5.5, the authors said, "cross-validation with the IDM". Cross-validation is a specific term. It would be better to use "comparison".

We thank the reviewer for pointing at the incorrect usage of the technical term "cross validation" and have revised the wording as suggested.

**References**

Yuan, Y., Zhang, Z., Yang, X. T., & Zhe, S. (2021). Macroscopic traffic flow modeling with physics regularized Gaussian process: A new insight into machine learning applications in transportation. Transportation Research Part B: Methodological, 146, 88-110.

Ahamed, T., Zou, B., Farazi, N. P., & Tulabandhula, T. (2021). Deep Reinforcement Learning for Crowdsourced Urban Delivery. Transportation Research Part B: Methodological, 152, 227-257.

Zhou, Y., & Ahn, S. (2019). Robust local and string stability for a decentralized car-following control strategy for connected automated vehicles. Transportation Research Part B: Methodological, 125, 175-196.

**Additional References**

Liu, C., Xu, X., & Hu, D. (2014). Multi-objective reinforcement learning: A comprehensive overview. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 45(3), 385-398.

Träuble, F., Dittadi, A., Wuthrich, M., Widmaier, F., Gehler, P. V., Winther, O., ... & Bauer, S. (2021). Representation Learning for Out-of-distribution Generalization in Reinforcement Learning. In ICML 2021 Workshop on Unsupervised Reinforcement Learning.