Dear reviewers:

I sincerely thank you for your careful review and valuable comments on this paper! The issues you have pointed out are extremely constructive and crucial to improving the quality of the paper. We have carefully revised the paper one by one, and now we would like to report the revisions as follows:

1.abstract：The abstract is overloaded with technical jargon, making it difficult to follow. Key terms such as “multi-agent deep reinforcement learning” and “proximal policy optimization” are introduced without sufficient context, which may confuse readers unfamiliar with the field. These concepts should be briefly defined, and their relevance to AV platoon merging should be explained. Additionally, the abstract is repetitive, using phrases like “optimization” and “efficiency” multiple times without clearly distinguishing their meaning in different contexts. Instead of broadly stating that the proposed method improves merging efficiency, energy consumption, and safety, the authors should specify by how much and in comparison to which baseline methods.

1)We have added explanations of terminology in the summary, the developed MA-DRL architecture enables coordinated learning among multiple autonomous agents to address the multi-objective coordination challenge through synchronized control of platoon longitudinal acceleration, AV steering and acceleration. ……which extends the Multi-agent PPO algorithm (a policy gradient method ensuring stable policy updates) by incorporating maximum Q-value action selection for platoon gap control and discrete command generation. As Section 3.3.1，second paragraph in blue.

2)Quantitative evaluation metrics: for the problem of new self-driving vehicles converging into an existing platoon, this paper may be the first to propose a unified framework, at present, the existing research mainly focuses on the one-sided (such as vehicle lateral control, trajectory planning, and longitudinal control of the fleet), therefore, we add a new comparative experiment of commonly used quintic polynomials+PID control+CACC for the dense traffic flow scenarios, and the experimental results show that this paper. The proposed method reduces the merging time by 37.69% (12.4s vs. 19.9s) and energy consumption by 58% (3.56kWh vs. 8.47kWh)." As Section 4.2，fourth and fifth paragraph in blue.

2. Introduction.

The first sentence could also be restructured to be more direct. For example, instead of stating, “With the advancement of autonomous driving and vehicle-to-everything (V2X) technologies, platooning has emerged as a frontier in intelligent transportation system research,” a more precise version would be: “Autonomous driving and vehicle-to-everything (V2X) technologies enable vehicle platooning, which enhances traffic efficiency and safety. However, integrating autonomous vehicles into existing platoons remains a

challenge due to coordination complexities.”

The introduction also lacks a clear structure. It jumps between concepts without a logical flow, making it difficult to follow. The transition from traditional methods to learning-based approaches should be smoother, explaining why classical control methods struggle with non-linearity before introducing deep reinforcement learning as a solution.

Literature Review.

I think it would be good to separate the introduction from the literature review.

The literature review presents many references but lacks a critical discussion of their limitations. Instead of listing previous studies, the authors should analyze their shortcomings and explicitly justify why the proposed approach is necessary. For example, several studies using PID, MPC, and reinforcement learning are cited, but their performance in similar merging scenarios is not compared. When referencing Gaagai et al.’s distributed controller and Tapli et al.’s cooperative adaptive cruise control (CCAC) algorithm, the weaknesses of these methods—such as computational complexity or instability under dynamic conditions—should be discussed. Additionally, some citations are outdated or do not seem directly relevant to the study. Instead of referencing general deep learning papers, the authors should link them specifically to the challenges of AV platoon merging. A table summarizing key differences between classical, optimization-based, and learning-based methods would improve readability and highlight the novelty of the proposed work.

1)We follow your comments and requests to adjust the introduction section accordingly and refine the literature review (1.1Platoon Longitudinal Control, 1.2Single-autonomous-vehicle merging control, 1.3 Single-autonomous-vehicle trajectory planning) In order to make the article logical, we have reorganised the language and added transitional sentences to make the logic clearer; we have also added a table (table1) to summarise the literature for easier reading.

2)We have summarized the limitations of the literature in the introduction section according to your suggestions and requirements and added tables to enhance readability.

3.Methodology (Problem formulation)

The methodology section lacks clarity in key areas. The problem formulation introduces a Markov Decision Process (MDP) but does not clearly define the state and action spaces. It is unclear how environmental uncertainty is handled—are stochastic disturbances considered? The description of the MAMQPPO algorithm is overly technical without sufficient explanation for a broader audience. When introducing the PD-Reward function, the authors state that it “reduces learning complexity and accelerates convergence,” but no theoretical justification or empirical evidence is provided. The reward function is described as “multi-

dimensional," but the explanation of how rewards are assigned and weighted remains vague. The weighting coefficients should be explicitly defined and justified based on experiments or previous literature.

The notation in Equations (1)–(4) is inconsistent, sometimes using subscripts and other times omitting them. For example $r=w_s\times r_s+w_c\times r_c$ and could be $r=w_{safety}\times r_{safety}+w_{comfort}\times r_{comfort}$

Additionally, there are multiple occurrences of "[Error! Reference source not found.]," indicating missing citations or figure references. These issues must be corrected before submission.

1)We have added a section on modeling Markov decision processes (see Section 2.1, second paragraph in blue). Additionally, this paper proposes an MAMQPPO algorithm, which is an improvement over the PPO algorithm. The PPO algorithm maintains policy exploration intrinsically through entropy regularization, with its core principle being the introduction of an entropy reward term in the optimization objective to dynamically adjust the flatness of the policy distribution. The PPO algorithm itself already has the capability to handle perturbations, detailed theoretical background can be found in the literature:[Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O., 2017. Proximal policy optimization algorithms. arxiv preprint arxiv:1707.06347.]

[Lee, H.K. and Yoon, S.W., Flat Reward in Policy Parameter Space Implies Robust Reinforcement Learning. In *The Thirteenth International Conference on Learning Representations*.]

[Zhang, H., Chen, H., **ao, C., Li, B., Liu, M., Boning, D. and Hsieh, C.J., 2020. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, *33*, pp.21024-21037.]

2)We have added a new paragraph describing the technical route to the MAMQPPO algorithm (see the fourth paragraph in blue of Section 3.3), along with a detailed description of the construction of the PD-Reward function in Section 2.3, first paragraph in blue.

3)We target the formula subscripts and the "[Error! Reference source not found.]", indicating missing citations or chart references, etc., are carefully modified to ensure proper submission formatting.

4)In the experimental phase, we revert to a detailed statement containing the lane setup vehicle position and initial conditions, etc., as detailed in (as section 4.1, first paragraph in blue)

5.Experimental Analysis

The experimental setup lacks critical details. The study mentions using a highway simulation

but does not specify the simulator or its validation process. How do the traffic conditions compare to real-world scenarios? A clearer explanation of lane configurations, vehicle positions, and initial conditions is needed. The choice of baseline algorithms (DDPG, TD3, SAC, PPO) is reasonable, but the authors do not justify why these were selected. Why were other multi-agent reinforcement learning approaches, such as MADDPG, not considered?

The evaluation metrics focus on merging time, success rate, and energy consumption, but computational efficiency is not analyzed. Given that MAMQPPO is a hierarchical approach, how does its training time compare to single-agent methods? The claim that "the AV safely merges into the platoon at a speed of 22.6m/s within 13.5s" needs further discussion—how do these values compare to human-driven vehicles or other control policies? Additionally, many figures are referenced incorrectly, making it difficult to follow the results. All figure captions should be descriptive, and trends in the data should be clearly explained.

1)The baseline reinforcement learning algorithm selected in this paper is more widely used in the current research field of intelligent transport, according to what you said about the MADDPG algorithm, because the DDPG algorithm is a global evaluation network, which can not meet the PD-Reward function designed in this paper.

2)We have reorganized the experimental analysis graphs and language, as shown in (in section 4.1, in blue font).

6.Conclusion and discussion (missing).

A discussion section is mandatory, and this manuscript lacks it.

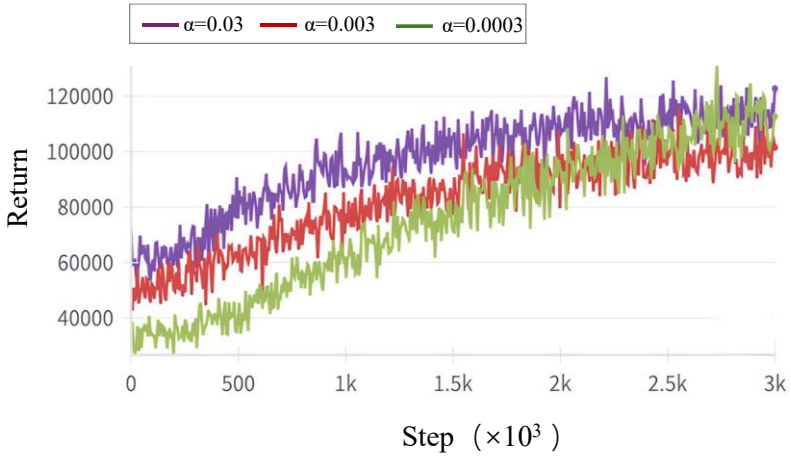
The conclusion overstates the impact of the proposed approach without sufficient supporting evidence. The authors claim that the method "effectively improves efficiency and reduces energy consumption while ensuring safety," but improvements are not quantified compared to existing methods. The discussion should be more nuanced, acknowledging potential limitations, such as high computational demands or sensitivity to hyperparameter tuning.

Additionally, the conclusion does not mention possible future research directions. Would integrating real-world vehicle data improve model generalization? How could transfer learning be applied to reduce the training burden? Ending with an open-ended discussion on these aspects would strengthen the study's contribution.

1)In response to your question, we have added an experiment with a fifth degree polynomial + PID + CACC from an existing method to compare with our proposed model and analyse the data in detail.(See Section 4.2, blue font section, for more details.).

2)We designed several sets of experiments for the learning rate α and found that the best

convergence was achieved when α = 0.03, so in this paper we set it to 0.03:



learning rate α: When α = 0.03, the model reaches peak returns (>1.2 × 10^5) in 1.5 × 10^3 steps, with the fastest convergence, but with ±8% late fluctuations; α = 0.003 requires 2 × 10^3 steps to stabilise to 1.0 × 10^5, with a 37% lower speed of convergence, but with better stability; and α =0.0003 is the least efficient, reaching only 8 × 10^4 in 3 × 10^3 steps. Experiments show that a high learning rate (α = 0.03) reinforces early exploration capability through fast gradient updating and is suitable for real-time demands of queue control (e.g., emergency lane changing), although stability risks need to be balanced, while a low learning rate (α = 0.003) is suitable for long-term strategy optimisation scenarios. In this paper, we choose α = 0.03 to prioritise the real-time control objective.

3).We have added new Future Directions content: The limitations of this study are threefold: first, model validation is based solely on simulated highway scenarios and does not cover the complex dynamic environment of urban traffic; second, multi-vehicle coordinated merging uses a phased serial strategy, lacking parallel decision optimization, which limits reorganization efficiency; third, it does not consider the interference of adverse weather conditions such as rain and snow on sensor . Future work will focus on: 1) integrating real platoon traffic data to enhance generalization capabilities in complex scenarios; 2) developing a multi-vehicle parallel merging mechanism based on game theory or swarm reinforcement learning to optimize global decision-making efficiency; 3) incorporating weather disturbance factors and robust control modules to improve model reliability in extreme environments.

7.Specific comments

The references do not follow the template of the journal

The manuscript contains several grammatical errors and awkward phrasing. For example, the sentence “the model achieves collaborative optimization of longitudinal platoon movements and AV lateral-longitudinal motions through integrated control” is difficult to parse and

should be rephrased for clarity. Some technical terms are used inconsistently, such as switching between “multi-agent deep reinforcement learning” and “multi-agent DRL.” Consistency should be maintained throughout the text.

Additionally, the figures are poorly formatted, with some missing labels or unclear axis descriptions. The authors should ensure that all figures are referenced in order and provide a detailed legend where necessary.

1)We made careful changes to information such as references and figure serial numbers to ensure proper submission formatting. At the same time, we adjusted the technical terminology of the paper accordingly, describing it with the full name for the first time and the abbreviation for the subsequent ones.

Thank you once again for your professional advice!

The above modifications have been verified through simulation experiments and updated in the manuscript, significantly enhancing the paper's completeness and practical value. If you have any other suggestions, we are willing to make further adjustments.