Trust-Based Multi-Agent Framework for On-Ramp Merging Integrating Large Language Models

Tianyi Wang*
The University of Texas at Austin
Austin, TX 78712, USA
bonny.wang@utexas.edu

Tianyi Zeng
Shanghai Jiao Tong University
Shanghai 200240, China
zengtianyi@sjtu.edu.cn

Yangyang Wang
Tongji University
Shanghai 201804, China
wyangyang@tongji.edu.cn

Junfeng Jiao
The University of Texas at Austin
Austin, TX 78712, USA
jjiao@austin.utexas.edu

Christian Claudel
The University of Texas at Austin
Austin, TX 78712, USA
christian.claudel@utexas.edu

Abstract

Intelligent transportation systems require connected and automated vehicles (CAVs) to conduct safe and efficient cooperation with human-driven vehicles (HDVs) in complex real-world traffic environments. However, the inherent unpredictability of human behaviors, especially at bottlenecks such as highway on-ramp merging areas, often disrupts traffic flow and compromises system performance when a deep understanding of the environment, the agents' intentions and humans' driving styles across various scenarios is needed. To address the challenge of cooperative on-ramp merging in heterogeneous traffic environments, this paper introduces a trust-based multi-agent (Trust-MA) framework, integrating trust-based reinforcement learning (RL) for individual interactions, a fine-tuned large language model (LLM) for regional cooperation, a reward function for global optimization, and the retrieval-augmented generation (RAG) mechanism to dynamically optimize decision-making across complex driving scenarios. Comparative experiments validate the effectiveness of the proposed Trust-MA approach, demonstrating significant improvements in safety, efficiency, comfort, and adaptability in multi-agent environments. These findings highlight the significant impact of cascading coordinated communication and dynamic functional alignment in advanced, human-like multi-agent autonomous driving environments.

1 Introduction

With the rapid development of connected and automated vehicles (CAVs), cooperative control based on CAVs has emerged as a solution to transform urban mobility and intelligent transportation systems [1–3]. Reinforcement learning (RL) stands out as a subfield that focuses on enabling agents to learn decision-making and action-taking in multi-agent environments, striving to achieve specific

*Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: UrbanAI: Harnessing Artificial Intelligence for Smart Cities.

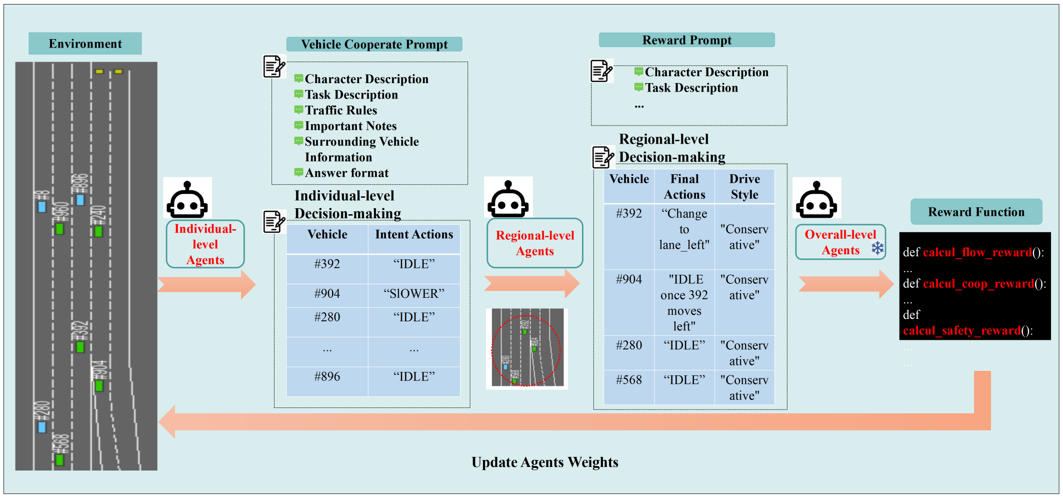

Figure 1: Overview of the proposed hierarchical multi-agent optimization framework [1]

goals through trial-and-error learning [4–7]. However, these approaches encounter several challenges: 1) Sample inefficiency and lack of generalization in online learning. 2) Difficulty in promoting general cooperative behaviors across diverse tasks. 3) Complexity in interpreting cooperative mechanisms.

Trust, as observed in human interactions, is a dynamic construct shaped by past interactions, behavioral expectations, and real-time observations [8]. In human-centered autonomous systems, trust has been widely studied as a means to enhance collaboration and mitigate conflict in shared environments [9, 10]. However, existing approaches often treat trust as a static or linearly evolving variable [11], and are largely confined to homogeneous agent settings [12], limiting their generalization and adaptability to real-world traffic scenarios.

Notably, devising accurate reward functions to encourage safe, efficient, and human-like driving behaviors remains a significant hurdle. A promising shift is emerging with the advent of in-context learning in multi-agent systems [13, 14], where large language models (LLMs) learn and make predictions based on immediate examples within the input context, without the need for additional fine-tuning or retraining. Through natural language prompts to specify objectives, LLMs translate agents' intentions into actionable reward signals, intuitively guiding RL agents towards desired outcomes [15]. This innovative interaction between LLMs and RL agents not only advances automated driving technology but also highlights LLMs' versatility in complex decision-making processes. However, it is challenging to combine RL agents and LLM, because the RL method requires a large number of interactive samples and is prone to local optimality. In contrast, the LLM enhancement method has strong generalization capabilities, but it is difficult to meet real-time response requirements.

To overcome these challenges, a novel trust-based multi-agent (Trust-MA) framework is proposed in this paper. The core idea of our framework is to leverage RL for individual-level interactions to ensure efficiency and responsiveness, while integrating fine-tuned LLM for smooth collaboration at the regional and global level. Besides, integrating a dynamic trust mechanism enables CAVs to continuously assess the reliability and behavioral tendencies of neighboring vehicles based on historical interactions and real-time observations. The main contributions of this paper are as follows:

- **Hierarchical Multi-Agent Optimization:** (Figure 1), which divides the multi-agent optimization problem into three levels: individual, regional, and global level.
- **Dynamic Trust Mechanism** (Figure 2), which is continuously evolved based on historical interactions and real-time observations to modulate cooperative behaviors.
- **Integration of RL and LLMs**, which adopts RL's high responsiveness for individual optimization, and a fine-tuned LLM to utilize its spatial and advanced semantic reasoning capabilities, augmented by road network images and high-level semantic information.

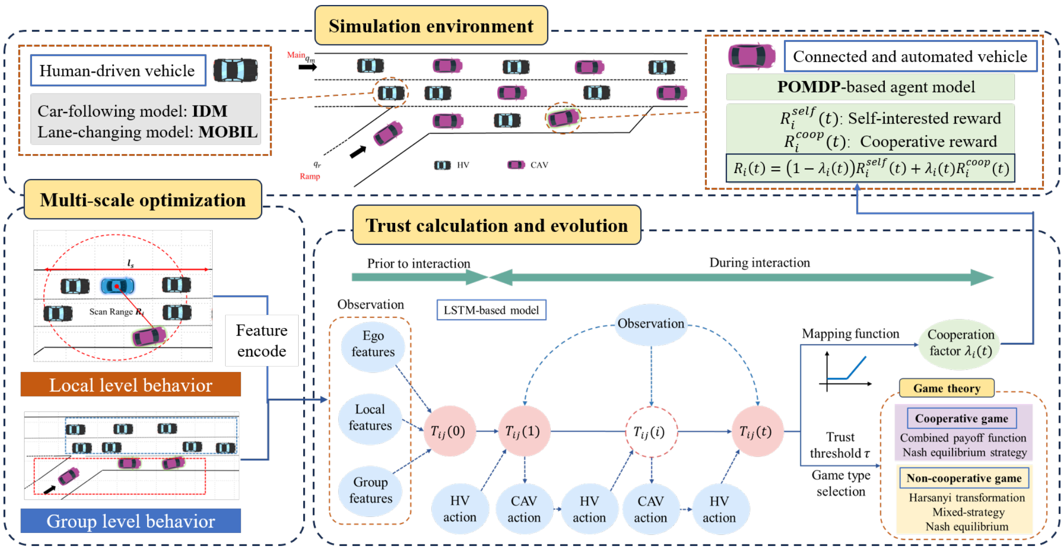

Figure 2: Overview of the proposed dynamic trust mechanism

2 Methodology

2.1 Partially Observable Markov Decision Process

CAVs are modeled as decentralized agents in a partially observable Markov decision process (POMDP), defined by the tuple:

$$\mathcal{M} = \langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}, \mathcal{P}, \{\mathcal{R}_i\}, \{\mathcal{O}_i\}, \gamma \rangle \tag{1}$$

where $\mathcal{N}$: The set of agents, $\mathcal{S}$: The global environment state, $\mathcal{A}_i$: The joint action space of agent i at each timestep, $\mathcal{P}$: The transition function governed by car-following and lane-changing dynamics, $\mathcal{R}_i$: The reward function for agent i, which is introduced in detail in the next subsection, $\mathcal{O}_i$: The observation space of agent i, which includes local features within its scan range, and γ: The discount factor for future rewards.

Each CAV agent $i \in \mathcal{N}_{\text{CAV}}$ receives an observation $o_i \in \mathcal{O}_i$ of the environment and selects an action $a_i \in \mathcal{A}_i$ based on a policy $\pi_i(o_i)$ to maximize expected discounted reward:

$$\mathcal{J}_i(\pi_i) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_i(s_t, a_t)\right] \tag{2}$$

2.2 Reward Function Design

Trust is introduced as a latent variable $T_{ij}(t)$, representing agent i's belief in agent j's cooperative intention at time t. This trust level modulates cooperation strategies through a weighting factor $\lambda_i(t)$:

$$\mathcal{R}_i(t) = (1 - \lambda_i(t))\mathcal{R}_i^{\text{self}}(t) + \lambda_i(t)\mathcal{R}_i^{\text{coop}}(t) \tag{3}$$

The self-interested reward $\mathcal{R}_i^{\text{self}}(t)$ for agent i is composed of three weighted components:

$$\mathcal{R}_i^{\text{self}}(t) = w_s \cdot r_i^{\text{safety}}(t) + w_c \cdot r_i^{\text{comfort}}(t) + w_e \cdot r_i^{\text{efficiency}}(t) \tag{4}$$

with:

$$r_i^{\text{safety}}(t) = -\mathbb{1}\left[d_{\mathrm{i}}^{headway}(t) < d_{\text{safe}}\right] \tag{5}$$

$$r_i^{\text{comfort}}(t) = -\left|\frac{da_i(t)}{dt}\right| \tag{6}$$

$$r_i^{\text{efficiency}}(t) = -\left|v_i(t) - v_{\text{desired}}\right| \tag{7}$$

where $d_{\text{i}}^{headway}(t)$ is the distance to the front vehicle, d_{safe} is the safety threshold, $\frac{da_i(t)}{dt}$ is the jerk (first derivative of acceleration), $v_i(t)$ is the current speed, v_{desired} is the desired free-flow speed, and $\mathbb{1}[x]$ is defined as a random variable which is 1 if x is true, and 0 if x is not true. The weights of these components (w_s, w_c, and w_e) are dynamically adjusted by the LLM based on current traffic conditions, leveraging our RAG mechanism.

The cooperative reward $\mathcal{R}_i^{\text{coop}}(t)$ is computed as follows:

$$\mathcal{R}_i^{\text{coop}}(t) = \frac{1}{Z_i(t)} \sum_{j \in \mathcal{N}_i(t)} T_{ij}(t) \cdot \mathcal{R}_j^{\text{self}}(t) \tag{8}$$

where $Z_i(t) = \sum_{j \in \mathcal{N}_i(t)} T_{ij}(t)$ is a normalization factor, and $\mathcal{N}_i(t)$ denotes the set of neighbors within the scan range. Then, we define a trust-based cooperation factor $\lambda_i(t)$, which determines the balance between self-interested and cooperative rewards in the total reward signal:

$$\lambda_i(t) = \min\left(1, \max\left(0, \frac{T_{ij}(t) - \tau}{1 - \tau}\right)\right), \tag{9}$$

where τ is a cooperation threshold. A higher trust value increases the influence of cooperative components in the reward function, encouraging socially aligned human behavior.

2.3 Dynamic Trust Mechanism

Let $T_{ij}(t) \in [0, 1]$ denote the trust level that agent i assigns to agent j at time step t. Trust is dynamically updated after each interaction using an exponentially weighted formulation:

$$T_{ij}(t+1) = (1 - \hat{\alpha}) T_{ij}(t) + \hat{\alpha} \cdot \delta(a_j^t), \tag{10}$$

where $\hat{\alpha} \in (0, 1)$ is a smoothing factor and $\delta(a_j^t) \in \{0, 1\}$ indicates whether agent j's behavior at time t is perceived as cooperative from agent i's perspective. Specifically:

$$\delta(a_j^t) = \begin{cases} 1, & \text{if } a_j^t \text{ contributes to agent } i \text{ achieving its goal,} \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

This update mechanism allows the trust signal to encode both short-term behavior and long-term cooperative trends, dynamically reflecting the reliability and social alignment of neighboring agents.

2.4 Large Language Model Coordination and Communication

At the individual level, RL agents process local state information from the environment, including vehicle states and traffic conditions, to generate initial actions. These actions are then evaluated through a policy mapping:

$$\pi(a_t^i | s_t^i) = \text{softmax}(\text{LLM}(s_t^i, \theta)), \tag{12}$$

where θ represents the LLM parameters and s_t^i encodes the agent's current state. The LLM evaluates this state information and produces logits that determine the likelihood of each possible action, ensuring responsive local decision-making. It then analyzes these inputs during the critical thinking phase to identify potential interactions and conflicts within its communication radius. In the final reflection phase, the LLM assigns driving styles (conservative or aggressive) to each vehicle using LoRA-based fine-tuning, adapting behavior to specific traffic contexts. The communication between agents is facilitated through a vector generated by the LLM:

$$m_t^i = \text{LLM}(s_t^i, a_t^i, \theta) \tag{13}$$

This vector m_t^i contains essential state and intent information, enabling coordinated decision-making among nearby vehicles. These coordinated decisions are then fed back to the global level, where they influence the reward generation process through the RAG system. The LLM continuously updates these communication vectors based on changing traffic conditions and vehicle interactions, creating a dynamic feedback loop that ensures smooth traffic flow through real-time adaptation of driving behaviors.

3 Experiments

3.1 Experiment Setup

Initially, simulation data are gathered, including screenshots from the HighWay Env simulator and vehicle details generated by a well-trained RL model. The dataset comprises approximately 50,000 samples, derived from six different road maps and simulated around 600 times each, with three different traffic densities and random spawn points, capturing about 10-20 timestamps per simulation. The fine-tuning process of GLM-4v-9B with LoRA [16] involves applying low-rank decomposition to the model's linear layers, where only the low-rank matrices are updated, thereby reducing memory and computational overhead. The LoRA layers of the model are fine-tuned on the dataset mentioned above in PyTorch, with a learning rate of 5×10^{-4}, a LoRA dimension of 16, and a LoRA alpha of 64, using two A100 GPUs, while the remaining parameters are frozen. The performance metrics considered include: **Merging Success Rate**, **Average Merging Time**, **Collision Rate**, and **Throughput**. We compare the proposed method with several SOTA MARL benchmarks, including the multi-agent advantage actor-critic (MAA2C) [17], the multi-agent proximal policy optimization (MAPPO) [18], and the multi-agent advantage actor-critic with Kronecker-factored trust region (MAACKTR) [19].

3.2 Experiment Results

Table 1: Performance under different traffic modes

(a) Easy Mode (4-7 vehicles in control range)

Model	Merging Success Rate (%)	Avg. Merging Time (s)	Collision rate	Throughput (vehicles/hour)
MAA2C	78 ± 2	25 ± 1.2	0.03± 0.01	271 ± 10
MAPPO	86± 2	24 ± 1.1	0± 0.01	290 ± 10
MAACKTR	75± 1.5	24 ± 1.1	0.09± 0.01	273 ± 10
Trust-MA	**92± 1.5**	**20 ± 1.0**	**0 ± 0.01**	**297± 10**

(b) Medium Mode (5-10 vehicles in control range)

Model	Merging Success Rate (%)	Avg. Merging Time (s)	Collision Rate	Throughput (vehicles/hour)
MAA2C	58 ± 2	28 ± 1.3	0.10 ± 0.02	487 ± 15
MAPPO	70 ± 2	27 ± 1.2	0.04 ± 0.01	563± 14
MAACKTR	60 ± 1.5	27 ± 1.2	0.14 ± 0.02	492± 15
Trust-MA	**83 ± 1.5**	**23 ± 1.1**	**0 ± 0.01**	**583 ± 12**

(c) Hard Mode (6-13 vehicles in control range)

Model	Merging Success Rate (%)	Avg. Merging Time (s)	Collision Rate	Throughput (vehicles/hour)
MAA2C	55 ± 2	35 ± 1.5	0.33 ± 0.15	780 ± 20
MAPPO	64 ± 2	34 ± 1.5	0.19 ± 0.1	810 ± 18
MAACKTR	57 ± 1.5	33 ± 1.5	0.26 ± 0.05	783± 19
Trust-MA	**75 ± 1.5**	**29 ± 1.2**	**0 ± 0.01**	**953 ± 15**

In low traffic density (Easy Mode) scenarios, our **Trust-MA** achieved a merging success rate of 0.92. The average merging time was 20 seconds, with no recorded accidents. These results indicate that under low traffic density conditions, the **Trust-MA** can perform merging maneuvers efficiently and safely with minimal delays. With medium traffic density (Medium Mode), the merging success rate slightly decreased to 0.83. The average merging time increased to 23 seconds, with no observed collisions in this capacity, highlighting areas for further optimization. The results suggest that while the system remains effective under moderate traffic, there is a slight increase in complexity and risk. The high traffic density (Hard Mode) scenarios presented significant challenges, but the system managed a 0.75 merging success rate. The average merging time was 29 seconds, with no accidents happened. These results indicate the need for improved coordination under high-density conditions.

4 Conclusion

This study presents the Trust-MA framework, aimed at enhancing autonomous driving behaviors through the integration of RL models and LLMs. Experimental results indicate that the Trust-MA framework significantly outperforms traditional algorithms, achieving higher success rates, lower collision rates, and enhanced traffic efficiency. In future work, we should refine the framework to improve its adaptability and robustness in managing a wider range of traffic scenarios, such as roundabouts and intersections, particularly in high-density conditions.

References

[1] M. Zhang, Z. Fang, T. Wang, S. Lu, X. Wang, and T. Shi, "Ccma: A framework for cascading cooperative multi-agent in autonomous driving merging using large language models," *Expert Systems with Applications*, vol. 282, p. 127717, 2025.

[2] Y. Wang and T. Wang, "Research on dual clutch intelligent vehicle infrastructure cooperative control based on system delay prediction of two lane highway on ramp merging area," *Automotive Innovation*, vol. 7, pp. 588–601, 2024.

[3] T. Wang, Q. Guo, C. He, H. Li, Y. Xu, Y. Wang, and J. Jiao, "Impact of connected and automated vehicles on longitudinal and lateral performance of heterogeneous traffic flow in shared autonomy on two-lane highways," *SAE Technical Paper*, no. 2025-01-8098, 2025.

[4] A. Khamis, *Optimization Algorithms: AI techniques for design, planning, and control problems*. Simon and Schuster, 2024.

[5] R. Sutton and A. Barto, "Reinforcement learning: An introduction," *Robotica*, vol. 17, no. 2, pp. 229–235, 1999. [Online]. Available: https://doi.org/10.1017/S0263574799271172

[6] B. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, 2022. [Online]. Available: https://doi.org/10.1109/TITS.2021.3054625

[7] W. Duan, Z. Tang, W. Liu, and H. Zhou, "Autonomous driving planning and decision making based on game theory and reinforcement learning," *Expert Systems*, vol. 40, no. 5, p. e13191, 2023. [Online]. Available: https://doi.org/10.1111/exsy.13191

[8] P. Hansen-Estruch, I. Kostrikov, M. Janner, J. G. Kuba, and S. Levine, "Idql: Implicit q-learning as an actor-critic method with diffusion policies," *arXiv preprint arXiv:2304.10573*, 2023.

[9] X. Wang, Z. Shi, F. Zhang, and Y. Wang, "Mutual trust based scheduling for (semi)autonomous multi-agent systems," in *2015 American Control Conference (ACC)*, 2015, pp. 459–464.

[10] C. Basu and M. Singhal, "Trust dynamics in human autonomous vehicle interaction: A review of trust models," in *Proceedings of the 2016 AAAI Spring Symposium Series*, 2016.

[11] Z. Zhou, G. Liu, and Y. Tang, "Multiagent reinforcement learning: Methods, trustworthiness, applications in intelligent vehicles, and challenges," *IEEE Transactions on Intelligent Vehicles*, pp. 1–23, 2024.

[12] J. Monteil, R. Billot, J. Sau, F. Armetta, S. Hassas, and N. E. El Faouzi, "Cooperative highway traffic: multi-agent modelling and robustness to local perturbations," in *Transportation Research Board 92nd Annual Meeting*, 2013, p. 24 p.

[13] N. Wies, Y. Levine, and A. Shashua, "The learnability of in-context learning," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 36 637–36 651.

[14] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey on in-context learning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processin*, 2024, pp. 1107–1128. [Online]. Available: https://aclanthology.org/2024.emnlp-main.64

[15] M. Kwon, S. Xie, K. Bullard, and D. Sadigh, "Reward design with language models," *arXiv preprint*, vol. arXiv:2303.00001, 2023. [Online]. Available: http://arxiv.org/abs/2303.00001

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[17] G. Papoudakis, F. Christianos, L. Schäfer, and S. Albrecht, "Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks," *arXiv preprint*, vol. arXiv:2006.07869, 2020. [Online]. Available: http://arxiv.org/abs/2006.07869

[18] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," *Advances in neural information processing systems*, vol. 35, pp. 24 611–24 624, 2022.

[19] Y. Wu, E. Mansimov, R. Grosse, S. Liao, and J. Ba, "Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We presented our contribution and scope of this paper clearly in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitation and future work are claimed in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper doesn't include any theoretical derivation and proof.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The results can be easily reproduced based on the technical details of this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The authors don't release the data and code, but the results can be reproduced based on the technical details of this paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup is clearly presented in this paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The statistical significance of the experiments is introduced in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are provided in experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper discusses the potential societal impacts in conclusion section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper doesn't have such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: The paper doesn't use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The authors don't release the data and code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper doesn't involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper doesn't involve human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method in this research doesn't involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.