LLM-PPO Driver: Improving Autonomous Driving via LLM-Guided Reward Shaping and Imitation Learning

Ahmad Mouri Zadeh Khaki
Mechanical Engineering Research Institute
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
ahmadmouri@kaist.ac.kr

Kyunghwan Choi
Cho Chun Shik Graduate School of Mobility
Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea
kh.choi@kaist.ac.kr

Abstract— **Proximal Policy Optimization (PPO) has shown promise for autonomous driving; however, it suffers from sparse rewards, slow convergence, and unsafe behaviors due to exploration without prior knowledge. These limitations are particularly critical in safety-sensitive driving scenarios, where failure events are rare but severe. To address this issue, we propose LLM-PPO Driver, a framework that enhances PPO-based motion planning by incorporating high-level semantic driving knowledge from a Large Language Model (LLM). The LLM does not participate in real-time decision-making; instead, it provides structured prior knowledge that is integrated through reward shaping and imitation learning. This lightweight and modular design eliminates deployment-time inference overhead while guiding policy learning toward safer and more efficient behaviors. Experiments in the Gym highway-v0 environment demonstrate consistent improvements in task success and safety over a baseline PPO agent, with imitation learning yielding the largest performance gain. These results highlight the effectiveness of leveraging LLM-based prior knowledge to mitigate unsafe exploration and improve learning efficiency in autonomous driving.**

Keywords—autonomous driving, imitation learning, large language model (LLM), proximal policy optimization, reward shaping

I. Introduction

Autonomous driving represents a complex and safety-critical sequential decision-making problem that requires robust perception, reasoning, and control under dynamic and uncertain environments [1]. Reinforcement learning (RL) has emerged as a promising paradigm for autonomous driving due to its ability to learn end-to-end control policies directly from interaction with the environment, reducing the need for manually engineered rules [2]. Among RL algorithms, Proximal Policy Optimization (PPO) [3] has gained widespread adoption owing to its stability, sample efficiency, and effectiveness in continuous control tasks. PPO-based approaches have demonstrated encouraging results in various driving scenarios [4].

Despite its advantages, PPO suffers from several limitations when applied to autonomous driving tasks such as achieving reliable and safe performance, particularly in long-horizon and safety-sensitive applications. A primary challenge lies in the design of reward functions, which are often sparse, delayed, or insufficiently informative to capture complex driving objectives such as safety, comfort, and efficiency [5]. As a result, PPO agents may exhibit slow convergence, unstable exploration, or unsafe behaviors during training. These issues are further exacerbated in autonomous driving environments, where critical events such as collisions or traffic violations occur infrequently but have severe consequences. Without explicit guidance or prior knowledge, PPO must rely solely on trial-and-error interactions, making it difficult to learn robust and human-like driving policies within practical training budgets.

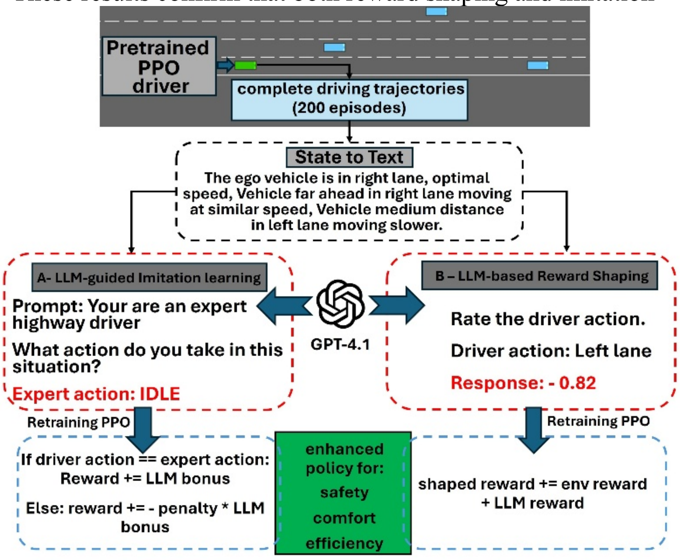
To address these challenges, incorporating prior knowledge into RL has been widely recognized as an effective strategy, commonly through techniques such as reward shaping and imitation learning [6, 7]. These methods can provide additional guidance to the agent, reduce exploration complexity, and encourage safer behaviors from early training stages. Recently, Large Language Models (LLMs) have emerged as powerful representations of human knowledge and reasoning, capturing rich semantic understanding of driving rules, traffic norms, and intuitive decision-making [8]. While LLMs have shown promise in high-level planning and decision support, their integration into low-level continuous control frameworks for real-time autonomous driving remains impractical due to prohibitive inference latency [9]. This gap motivates the exploration of LLMs as high-level knowledge providers that can guide policy optimization, rather than replacing established RL algorithms as the primary agents interacting with the environment.

In this work, we propose LLM-PPO Driver, a framework that enhances PPO for autonomous driving by leveraging an LLM as a source of high-level driving knowledge. Rather than replacing the underlying RL algorithm, the LLM is used as an auxiliary guidance module in two independent settings. In the first setting, the LLM is utilized for reward shaping, where high-level semantic driving knowledge is translated into informative reward signals that densify sparse feedback and encourage safety-aware behavior by rating agent actions. In the second setting, the LLM provides expert-like demonstrations that are incorporated through imitation learning to guide policy optimization toward human-consistent driving strategies. Each approach is evaluated separately to isolate its impact on PPO performance. This modular design preserves the stability of PPO while enabling systematic analysis of how LLM-driven prior knowledge improves autonomous driving policies. Fig. 1 illustrates the pipeline of our proposed design.

Experimental results demonstrate that incorporating LLM-driven expert knowledge significantly improves PPO performance in autonomous driving tasks. Compared to a baseline PPO agent achieving a mean success step (SS_{mean}) of 25.16 and a success rate (SR) of 70%, the LLM-based reward shaping approach yields notable gains, increasing SS_{mean} to 27.22 and SR to 76%. Similarly, the imitation learning variant guided by LLM-generated demonstrations further improves

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS2025-00554087).

performance, achieving an SS_{mean} of 27.85 and an SR of 82%. These results confirm that both reward shaping and imitation learning independently enhance policy optimization, with imitation learning providing the largest overall improvement. The key contributions of this work are threefold:



Pretrained PPO driver
complete driving trajectories (200 episodes)
State to Text
The ego vehicle is in right lane, optimal speed, Vehicle far ahead in right lane moving at similar speed, Vehicle medium distance in left lane moving slower.
A- LLM-guided Imitation learning
Prompt: Your are an expert highway driver
What action do you take in this situation?
Expert action: IDLE
GPT-4.1
B – LLM-based Reward Shaping
Rate the driver action.
Driver action: Left lane
Response: - 0.82
Retraining PPO
Retraining PPO
If driver action == expert action: Reward += LLM bonus
Else: reward += - penalty * LLM bonus
enhanced policy for:
safety
comfort
efficiency
shaped reward += env reward + LLM reward

Fig. 1. Overview of the LLM-PPO Driver framework.

- Introducing LLM-PPO Driver, a modular framework that systematically incorporates LLMs as high-level knowledge providers within PPO, enabling human driving intuition in low-level continuous control without compromising training stability or real-time execution.
- Designing two independent LLM-guided mechanisms—semantic reward shaping via action evaluation and imitation learning via expert demonstration—integrated separately into PPO. This separation enables clear performance attribution and insights into how LLM-derived prior knowledge influences policy learning in safety-critical driving.
- Validating consistent safety and task success improvements through extensive simulation. Results demonstrate that both LLM-driven strategies improve task completion and safety metrics, highlighting LLMs' effectiveness as auxiliary reasoning modules in autonomous driving RL.

The remainder of this paper is organized as follows. Section II reviews related work on utilizing RL methods, LLMs and integrating them for autonomous driving tasks. Section III provides the proposed methodology of LLM-PPO driver in detail. The results are presented and discussed in Section IV. Finally, Section V concludes the paper.

II. Related Work

A. RL-based Planning in Autonomous Driving

Autonomous driving systems typically use modular architectures with perception, prediction, and planning components. The planning module generates trajectories balancing safety, comfort, efficiency, and route progress. Well-defined module interfaces enable focused optimization and integration of learning-based methods.

RL is widely explored for driving planning, formulating it as sequential decision-making under dynamic traffic [11, 12]. Recent advances use deep neural networks and continuous control for complex scenarios [13]. PPO is particularly popular for its training stability and effectiveness in continuous action spaces [14].

However, RL-based planners remain sensitive to reward design and exploration, often suffering from sparse feedback and unsafe training behaviors. Prior works use heuristic reward shaping or expert demonstrations, but these rely on manually engineered rules and domain-specific heuristics. These limitations motivate exploring alternative prior knowledge sources that can guide RL-based planning more flexible and scalable.

B. Employing LLM in Autonomous Driving

Recent progress in autonomous driving has improved planning and decision-making [15], but learning-based systems still face challenges in data efficiency, interpretability, and incorporating human driving knowledge. LLMs have attracted attention due to their reasoning capabilities and ability to encode semantic knowledge, including traffic rules, conventions, and human decision-making [16, 17].

Most existing works employ LLMs as high-level planners or decision-makers through prompt engineering [18] or fine-tuning [19, 20] to generate driving commands, symbolic plans, or textual explanations. While these improve interpretability and reasoning, deploying LLMs in closed-loop control remains impractical due to inference latency, reliability concerns, and difficulty translating language outputs into precise, real-time control actions.

Rather than using the LLM as a planner or controller, we leverage LLMs as auxiliary knowledge providers that guide RL-based planning indirectly. By generating reward shaping signals and expert-like demonstrations, we transfer high-level semantic driving knowledge into PPO-based policy optimization without introducing LLMs into the real-time control loop. This preserves training stability and enables systematic evaluation of how LLM-derived guidance improves autonomous driving performance.

C. Integrating LLM and RL for Autonomous Driving

Contemporary efforts integrating LLMs with RL for autonomous driving have shown that LLMs can incorporate high-level semantic reasoning and human-like driving knowledge into data-driven decision-making. While these works demonstrate that LLMs can improve generalization, robustness, and interpretability in complex driving scenarios, they exhibit several limitations.

For example, a hierarchical framework where LLMs act as high-level planners generating goals, strategies, or meta-actions during execution has been proposed in [18]. However, this approach embeds the LLM directly into the decision-making pipeline. This tight coupling introduces sensitivity to LLM hallucinations and complicates real-time deployment. Alternatively, recent hint-based frameworks [19] reduce the direct influence of LLMs by treating their outputs as auxiliary signals, but still integrate LLM guidance into policy learning through state augmentation and additional control modules, increasing system complexity and runtime overhead. Teacher-student models that fuse LLM guidance with RL policies [20] and data-centric distillation pipeline methods [21] mitigate inference latency through offline LLM use. However, these approaches require extensive offline data generation, multi-policy architectures, and complex adaptation mechanisms, making performance attribution and scalability challenging.

In contrast, our methodology adopts a lightweight, modular strategy preserving the standard PPO pipeline. We use LLMs exclusively as offline knowledge providers to generate reward shaping signals or expert demonstrations, incorporated independently into training. LLMs are not involved in real-time decision-making or state augmentation, eliminating inference latency and reducing hallucination vulnerability. By avoiding hierarchical control, distillation, and auxiliary modules, our approach enables clear performance attribution, maintains training stability, and provides a practical pathway for incorporating semantic driving knowledge into RL-based planning.

TABLE I. HYPERPARAMETERS USED FOR TRAINING

Parameter	Value
Learning rate	3×10^{-4}
Batch size	64
Clip range	0.2
Discount factor	0.99
GAE factor	0.95

III. LLM-PPO DRIVER

A. *Baseline PPO and Simulation Environment*

In this work, we employ the PPO algorithm from the Stable Baselines3 library [25]. The model utilizes a multi-layer perceptron (MLP) policy architecture consisting of two hidden layers with 64 neurons each and Tanh activation functions. The specific hyperparameters used for training are summarized in Table I. We conduct experiments in the gym "highway-v0" environment [26]. The highway-v0 environment models 4-lane highway driving with surrounding traffic vehicles governed by rule-based controllers. At each time step, the agent receives a structured numerical observation (s_t) describing the driving scene. The state space (s) consists of normalized relative kinematic features of the ego vehicle and surrounding vehicles, including longitudinal and lateral positions and velocities (x, y, v_x, v_y).

The action space (a) is discrete, consisting of five high-level driving commands: lane change left, lane change right, maintain current speed and lane, accelerate, and decelerate. The PPO policy outputs a categorical distribution over these actions at each time step (a_t), enabling the agent to learn strategic lane-changing and speed-control behaviors under dynamic traffic conditions.

The reward function (r_t) follows the default formulation of the highway-v0 environment and combines safety and efficiency objectives. A collision incurs a terminal penalty of −1, enforcing strong safety constraints. To encourage efficient driving, the agent receives a right-lane reward of 0.1 and a high-speed reward of 0.4 when maintaining a velocity within the target range [20,30] m/s. Lane-change actions are assigned a neutral reward (0) to avoid explicitly biasing lateral maneuvers. All rewards are normalized, ensuring stable gradient updates during PPO training. This reward design provides only coarse guidance and remains insufficient to encode nuanced driving semantics, motivating the introduction of LLM-guided reward shaping and imitation learning.

B. *Overview of the Proposed Framework*

The proposed LLM-PPO Driver framework builds upon a standard PPO-based autonomous driving agent that operates using low-level continuous control actions (steering and acceleration). An LLM is incorporated as an expert driver that provides high-level semantic guidance during training only. Crucially, the LLM is never involved in online policy inference or real-time control, ensuring that the deployed agent remains lightweight, efficient, and compatible with real-world autonomous driving requirements. This design preserves the stability and scalability of PPO while enabling the transfer of human-like driving knowledge into policy learning.

First, a baseline PPO agent is trained using environment rewards alone. This baseline policy is then used to collect a fixed set of complete driving trajectories (200 episodes), each spanning from an initial state to termination. The collected data includes a set of $D = \{s_t, a_t, r_t, c, \mathit{info}\}$ for each time step, where c indicates whether the ego vehicle crashed and *info* contains information about the ego vehicle's speed and individual component of r_t. To enable LLM interaction, the agent's numerical observations are translated into textual descriptions of the driving scenario. The LLM is subsequently queried in two independent training pipelines: a) for imitation learning, where the LLM selects expert actions given the current state; or b) for reward shaping, where the LLM evaluates the baseline agent's actions and assigns semantic feedback scores. The resulting LLM-guided data are then used to retrain PPO, producing policies that better align with safety, efficiency, and human driving conventions—without modifying the PPO architecture or introducing additional control modules. Each model configuration was trained for a minimum of $5{\times}10^5$ timesteps, utilizing eight parallel environments to optimize computational throughput. For the expert supervision component of the framework, GPT-4.1 was employed as the primary LLM.

C. *Converting State to Text*

To enable the LLM to interpret driving scenarios and provide high-level semantic guidance, the numerical state observations from the highway-v0 environment are converted into structured natural language descriptions. Since LLMs operate on textual inputs, this conversion serves as a critical interface between low-level RL representations and high-level expert reasoning.

The kinematic observation of each time step (s_t) is mapped to a textual description (τ_t) that summarizes the ego vehicle's driving context and surrounding traffic. The ego vehicle is described in terms of its lane position (e.g., left, center, right) and discretized speed category (e.g., slow, moderate, optimal). Nearby vehicles are described relative to the ego vehicle using qualitative distance buckets (e.g., very close, close, far ahead/behind), relative lane positions, and speed relations (faster, slower, or similar speed). Only vehicles present in the observation are included, and the description is dynamically updated at each time step. Moreover, a set of $M = \{a_t, r_t, c, \mathit{info}\}$ is processed at each time step for reward shaping.

This abstraction transforms continuous and normalized numerical features into semantically meaningful and compact descriptions while preserving critical spatial and dynamic relationships. The resulting text representation provides sufficient contextual information for the LLM to reason about safety, efficiency, and driving intent, without exposing low-level state details. These textual descriptions are subsequently used as inputs to the LLM for both reward shaping and imitation learning, enabling the transfer of expert-level driving knowledge into PPO-based policy optimization.

D. *LLM-based Imitation Learning*

In this strategy, the LLM serves as an expert driving advisor that provides policy-level guidance for imitation learning. Given a textual description of the current driving state s_t produced by the state-to-text conversion module τ_t, the LLM evaluates all discrete control actions a and assigns relative preference scores $f_{LLM}(s_t,a) \in [-1,1]$, reflecting expert judgment in terms of safety, efficiency, and driving quality. This evaluation relies solely on the present observation, without access to environment rewards, future outcomes, or trajectory information, ensuring that the resulting guidance represents pure expert priors rather than reinforcement signals.

To obtain a stable and unambiguous supervision target (a_t^{expert}), the LLM-assigned preference scores $f_{LLM}(s_t,a)$ are converted into a discrete expert action via maximum-preference selection, forming a hard-matching supervision signal for imitation learning that subsequently guides PPO policy optimization, as expressed in (1).

$$a_t^{expert} = \arg\max_{a} \; f_{LLM}(s_t,a), \tag{1}$$

This hard selection strategy ensures that expert guidance remains consistent, interpretable, and robust to noisy or ambiguous preference signals. Consequently, semantic driving knowledge from the LLM can be transferred into the RL policy while preserving the independence of the learning process and avoiding direct LLM involvement in real-time control.

Expert actions generated by the LLM are collected offline and stored in a lookup table that maps normalized textual state descriptions to discrete expert actions. During training, a custom vectorized environment wrapper retrieves the expert action associated with the agent's previous observation and compares it with the action executed by the PPO policy, producing a deterministic imitation signal incorporated directly into the reward. When the agent's action matches the expert recommendation, a positive imitation bonus is applied; otherwise, a mismatch penalty is imposed. Observations without available expert annotations are ignored, allowing uninterrupted learning from environment interaction.

An effective training strategy is essential to fully exploit limited expert supervision while preserving the stability of on-policy RL. Accordingly, the agent learns simultaneously from environment rewards and LLM-derived guidance, enabling the policy to incorporate semantic driving knowledge without hindering exploration or convergence. The proposed wrapper operates transparently on top of PPO without modifying the policy architecture, observation space, or action space, thereby supporting efficient reuse of offline expert data, compatibility with vectorized training, and continuous monitoring of expert coverage, match rates, and imitation effectiveness. The final reward used for policy optimization is defined as:

$$r_t = r_t^{env} + \beta\, \mathbb{I}(at = a_t^{expert}) - \beta\alpha\, \mathbb{I}(at \neq a_t^{expert}), \tag{2}$$

where r_t^{env} denotes the original environment reward, a_t^{expert} the expert action provided by the LLM when available, β the imitation strength, $\mathbb{I}[\cdot]$ the indicator function, and α the mismatch-penalty coefficient. The imitation weight β follows an exponential decay schedule as shown in (3) defined by:

$$\beta = \max\,(\beta_{min}, \beta_0\, e^{-\kappa t}), \tag{3}$$

where β_0 is the initial weight (0.3), β_{min} is the lower asymptote (0.05), and κ is the decay constant (0.00001). This formulation allows for a stronger reliance on expert supervision during early training steps t and facilitates a gradual transition toward environment-driven optimization as the agent matures.

This joint learning mechanism enables the agent to benefit from both semantic expert guidance and experiential feedback, improving sample efficiency, behavioral consistency, and training stability while preserving the exploratory capacity of PPO.

E. Employing LLM for Reward Shaping

In this setup, the LLM is employed as a semantic reward evaluator that provides auxiliary feedback for reward shaping. Unlike the expert-action advisor described in Section III-D, which offers policy-level supervision prior to action execution, this mechanism operates post-action by assessing the quality of the agent's realized behavior using all available contextual information.

In contrast to the imitation-learning setting, the LLM receives a complete structured textual transition representation ($\tau_t + M$) at each step, including the observation, executed action, environment reward, and auxiliary metadata. Based on this comprehensive context, the LLM outputs a continuous quality score within the bounded interval $[-1,1]$, where negative values denote unsafe or poor decisions, values near zero indicate neutral or mediocre behavior, and positive values correspond to safe, efficient, and contextually appropriate driving. This bounded normalization facilitates stable integration of semantic feedback into the RL process.

To ensure computational efficiency, LLM evaluations are pre-computed offline and cached using a composite key formed from the observation–action pair (and associated transition context), enabling reuse of semantic scores for repeated transitions.

To integrate this semantic feedback into PPO in a stable and efficient manner, we formalize LLM-guided reward shaping as an additive modification of the environment reward, enabling continuous supervision without altering the policy representation or interaction dynamics. Formally, the final reward used for policy optimization is defined as:

$$r_t = r_t^{env} + \gamma\, r_t^{LLM}, \tag{4}$$

where r_t^{env} denotes the original environment reward, r_t^{LLM} represents the semantic quality score assigned by the LLM to the executed state–action pair, and γ is a fixed scaling coefficient controlling the contribution of LLM-derived feedback. Unlike the imitation-learning formulation in in Section III-D, no temporal decay is applied to β, ensuring consistent semantic guidance throughout training.

To operationalize this formulation, we implement a vectorized reward-shaping wrapper that augments the reward stream without modifying PPO's learning dynamics. At each transition, the wrapper converts the previous observation and executed discrete action into normalized textual representations via the state-to-text converter, forming a composite lookup key used to retrieve a pre-computed LLM reward from an offline cache. When a matching entry exists, the retrieved semantic score is scaled by β and additively combined with the environment reward. Otherwise, the LLM contribution defaults to zero, allowing uninterrupted learning from the environment signal alone. The wrapper additionally records cache hit and miss statistics, enabling quantitative measurement of semantic reward coverage during learning.

Overall, this formulation allows the LLM to function as a high-level semantic critic that complements the handcrafted environment reward. Whereas the environment reward encodes task-specific numerical heuristics, the LLM evaluation captures holistic behavioral quality by jointly considering safety, efficiency, situational appropriateness, and realized outcomes, thereby providing richer supervisory signals for policy optimization.

TABLE II. PERFROMANCE EVALUATION OF THE LLM-PPO DRIVER

Metrics	**Models**		
	Baseline	***Reward Shaping (γ=0.3)***	***Imitation Learning (β=0.3, α=0.2)***
SS$_{min}$	2.00	6.00	8.00
SS$_{Q1}$	23.75	30.00	30.00
SS$_{median}$	30.00	30.00	30.00
SS$_{Q3}$	30.00	30.00	30.00
SS$_{max}$	30.00	30.00	30.00
SS$_{mean}$	25.16	27.22	27.85
SR%	70.00	76.00	82.00

IV. RESULTS AND DISCUSSION

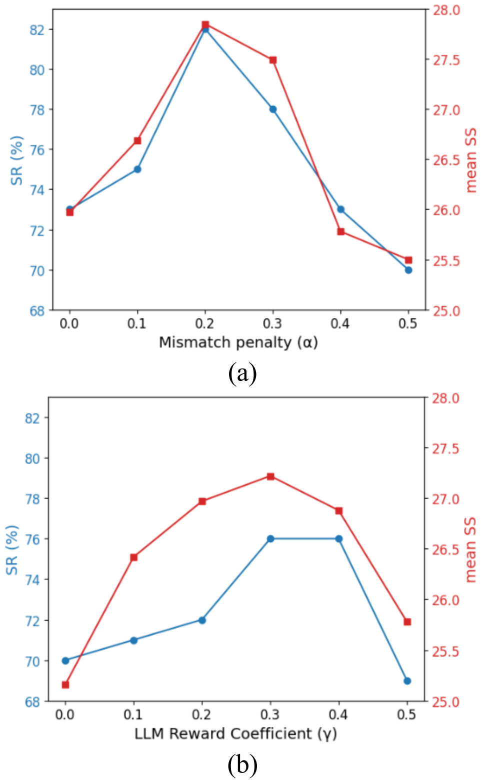
A. Evaluation metrics

To evaluate driving safety and task completion, we use two metrics: Survival Steps (SS) and Success Rate (SR). Each policy is evaluated over 100 episodes with deterministic actions, maximum length T = 30 steps, or earlier termination upon collision or completion. SS measures consecutive time steps without crashing. We report SS statistics across episodes: minimum (SSmin), first quartile (SSQ1), median (SSmedian), third quartile (SSQ3), maximum (SSmax), and mean (SSmean). SR is the proportion of episodes where the agent completes the full horizon without collision (SS = T). SS captures graded safety duration while SR reflects strict collision-free completion, providing comprehensive assessment of policy robustness. During evaluation, traffic density increases from 1.0 to 1.5 and vehicles from 50 to 70, creating denser scenarios for stricter safety assessment.

B. Experimental Results

Table II compares the baseline PPO agent with the proposed LLM-guided reward shaping and imitation learning strategies. Both LLM-based methods consistently improve safety and task completion over the baseline, with imitation learning achieving the strongest performance. In terms of survival behavior, the baseline exhibits poor worst-case robustness (SS$_{min}$ = 2), whereas reward shaping and imitation learning increase SS$_{min}$ to 6 and 8, respectively, indicating substantial reduction of early-collision episodes. Quartile statistics further show that both LLM-guided approaches achieve SS$_{Q1}$ = SS$_{median}$ = SS$_{Q3}$ = 30, meaning that at least 75 % of evaluation episodes survive the full horizon, while the baseline (SS$_{Q1}$ = 23.75) remains more failure-prone. Mean survival duration improves from 25.16 (baseline) to 27.22 with reward shaping and 27.85 with imitation learning. A similar monotonic trend is observed in collision-free completion, where SR increases to 76% and 82%, corresponding to relative improvements of 8.57% and 17.14% over the 70% baseline. These results demonstrate that offline LLM-derived semantic supervision significantly enhances PPO safety and robustness in dense highway scenarios, with direct policy-level imitation providing greater behavioral alignment and performance gains than post-hoc reward shaping.

The ablation results in Fig. 2 demonstrate that both the *α* in imitation learning and the γ in reward shaping exhibit a clear optimal operating region. Moderate increases in these coefficients initially improve performance, as reflected by higher success rates and longer survival durations, indicating that stronger semantic guidance enhances behavioral alignment and safety. However, further increases beyond the optimal values lead to performance degradation in both metrics. This decline suggests that excessive reliance on LLM-derived supervision can suppress effective exploration, thereby reducing adaptability to diverse traffic scenarios. These findings highlight the importance of balanced integration between environment-driven reinforcement and semantic LLM guidance, confirming that LLM assistance is most effective when used as a complementary signal rather than a dominant training objective.



(a)

(b)

Fig. 2. Ablation analysis of semantic guidance strength: (a) α in imitation learning and (b) γ in reward shaping

C. Discussion

The experimental results demonstrate that lightweight semantic integration of LLM guidance into PPO can substantially enhance autonomous driving performance without introducing architectural complexity or real-time inference overhead. Both imitation learning and reward shaping consistently improve survival duration and collision-free completion compared with the baseline, confirming that high-level semantic knowledge from LLMs can effectively complement environment-driven reinforcement signals. Critically, these gains are achieved through an offline and modular framework that preserves PPO stability, maintains computational efficiency, and enables clear attribution of performance improvements. These results establish the practical viability of incorporating LLM-derived semantic reasoning into safety-critical autonomous driving policies while preserving real-time control capabilities suitable for deployment.

The two guidance mechanisms exhibit complementary performance characteristics. Imitation learning delivers the highest safety gains, indicating stronger worst-case robustness. Reward shaping achieves more moderate improvements but shows smoother and more stable sensitivity to hyperparameters. Ablation results reveal that imitation learning is strongly dependent on the mismatch penalty α, with performance peaking near $\alpha = 0.2$ before degrading, whereas reward shaping varies more gradually with the reward weight γ and remains effective across a wider range. This contrast highlights a key trade-off: imitation learning provides higher peak performance through direct action supervision, while reward shaping offers greater robustness and stability under imperfect tuning. Consequently, imitation learning primarily accelerates convergence toward safe driving strategies, whereas reward shaping promotes gradual and stable performance improvement. Together, these findings indicate that policy-level supervision and reward-level semantic evaluation address distinct aspects of RL and may be synergistically combined to further enhance stability, safety, and generalization in autonomous driving.

Future work can extend this framework by exploring tighter integration between imitation learning and reward shaping, such as iterative or alternating training schemes that progressively refine both policy alignment and semantic reward feedback. Such hybrid strategies may further improve stability, safety, and sample efficiency. Another promising direction is the design of more elaborate and multi-objective reward functions that better balance safety, efficiency, and driving comfort to enable more effective and realistic learning. In addition, incorporating vision–language models (VLMs) capable of interpreting sensor-level observations, and validating the approach in large-scale realistic simulators such as nuPlan [27], would increase environmental fidelity and support the transition toward real-world autonomous driving deployment.

V. Conclusion

This paper presents a lightweight framework for integrating LLM-derived semantic knowledge into PPO-based autonomous driving without modifying the policy architecture or requiring real-time LLM inference. By leveraging offline LLM guidance through imitation learning and reward shaping, the approach preserves training stability and computational efficiency while transferring high-level driving semantics into RL. Experiments in dense highway scenarios demonstrated consistent improvements in survival duration and collision-free completion, confirming the effectiveness of semantically informed supervision for safety-critical decision making. These findings establish a practical pathway for incorporating foundation-model reasoning into deployable autonomous driving systems and motivate future work on unified LLM-RL integration and evaluation in large-scale realistic simulators.

References

[1] L. Chen et al., “Milestones in Autonomous Driving and Intelligent Vehicles: Survey of Surveys,” in IEEE Transactions on Intelligent Vehicles, vol. 8, no. 2, pp. 1046-1056, Feb. 2023.

[2] S. Aradi, “Survey of Deep Reinforcement Learning for Motion Planning of Autonomous Vehicles,” in IEEE Transactions on Intelligent Transportation Sys., vol. 23, no. 2, pp. 740-759, Feb. 2022.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, “Proximal Policy Optimization Algorithms,” arXiv preprint arXiv:1707.06347v2 [cs.LG], 2017.

[4] A. Ray, J. Achiam, and D. Amodei, “Benchmarking safe exploration in deep reinforcement learning,” 2019, OpenAI Blog. [online]. Available: https://openai.com/index/benchmarking-safe-exploration-in-deep-reinforcement-learning/

[5] B. R. Kiran et al., “Deep Reinforcement Learning for Autonomous Driving: A Survey,” in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 6, pp. 4909-4926, June 2022.

[6] M. Hallgarten, M. Stoll and A. Zell, “From Prediction to Planning With Goal Conditioned Lane Graph Traversals,” 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 2023, pp. 951-958.

[7] N. Di Paolo et al., “Towards A Unified Agent with Foundation Models,” arXiv preprint arXiv:2307.09668 [cs.RO], 2023.

[8] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi and Y. Qiao, “Drive like a human: Rethinking autonomous driving with large language models,” In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 910–919, 2024.

[9] S. Hu, Z. Fang, Z. Fang, X. Chen, Y. Fang, “AgentsCoDriver: Large Language Model Empowered Collaborative Driving with Lifelong Learning,” arXiv preprint arXiv:2404.06345 [cs.AI], 2024.

[10] X. Xu, L. Zuo, X. Li, L. Qian, J. Ren and Z. Sun, “A Reinforcement Learning Approach to Autonomous Decision Making of Intelligent Vehicles on Highways,” in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 50, no. 10, pp. 3884-3897, Oct. 2020.

[11] O. Scheel, L. Bergamini, M. Wolczyk, B. Osiński, and P. On druska, “Urban driver: Learning to drive from real-world demonstrations using policy gradients,” Proceedings of the 5th Conference on Robot Learning, PMLR, pp. 718-728, 2022.

[12] B. Kim, K. Kim, S. Kim, Y. Shin and H. Ahn, “Domain-Agnostic Scalable AI Safety Ensuring Framework,” arXiv preprint arXiv:2504.20924 [cs.AI], 2025.

[13] J. Kim, K. Choi, “CAR Planner: Constrained-Attention-Based Robust Imitation Learning for Autonomous Driving,” TechRxiv, 2025.

[14] Z. Tang et al., “VLMPlanner: Integrating Visual Language Models with Motion Planning,” arXiv preprint arXiv:2507.20342 [cs.AI], 2025.

[15] J. Cheng, Y. Chen, X. Mei, B. Yang, B. Li and Ming Liu, “Rethinking imitation-based planners for autonomous driving,” In 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 14123–14130, 2024.

[16] L. Wen et al., “DiLu: A Knowledge-Driven Approach to Autonomous Driving with Large Language Models,” arXiv preprint, arXiv:2309.16292 [cs], 2023.

[17] Z. Xu et al., “DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model,” in IEEE Robotics and Automation Letters, vol. 9, no. 10, pp. 8186-8193, Oct. 2024.

[18] L. Li, R. Tan, J. Fang, J. Xue and C. Lv, “LLM-augmented hierarchical reinforcement learning for human-like decision-making of autonomous driving,” Expert Sys. with Applications, vol. 294, 128736, Dec. 2025.

[19] Z. Chen et al., “HCRMP: ALLM-Hinted Contextual Reinforcement Learning Framework for Autonomous Driving,” arXiv preprint arXiv:2505.15793 [cs.RO], 2025.

[20] C. Xu et al., “TeLL-Drive: Enhancing Autonomous Driving with Teacher LLM-Guided Deep Reinforcement Learning,” arXiv preprint arXiv:2502.01387v3 [cs.AI], 2025.

[21] S. Wu et al., “Robust RL with LLM-Driven Data Synthesis and Policy Adaptation for Autonomous Driving,” arXiv preprint arXiv:2410.12568v2 [cs.RO], 2024.

[22] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus and N. Dormann, “Stable-Baselines3: Reliable Reinforcement Learning Implementations,” Journal of Machine Learning Research, vol. 22, no. 268, pp. 1-8, Nov. 2021.

[23] E. Leurent, “An environment for autonomous driving decision-making,” 2018, gitHub repository. [online]. Available: https://github.com/eleurent/ highway-env

[24] H. Caesar, J. Kabzan, K. S. Tan, W. K. Fong, E. Wolff, A. Lang, L. Fletcher, O. Beijbom, and S. Omari, “nuplan: A closed-loop ml based planning benchmark for autonomous vehicles,” arXiv preprint arXiv:2106.11810, 2021.