ENHANCED SIMULATION ENVIRONMENT BASED ON DISTRIBUTION MATCHING AND LARGE LANGUAGE MODEL-GUIDED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Autonomous driving simulation environments often suffer from systematic statistical drift during long-term closed-loop operation, where the behavior of background vehicles gradually deviates from real-world traffic statistics. Existing enhancement methods either rely on rule-based models with limited statistical fidelity or retain online large language model (LLM) calibration during simulation, which introduces prohibitive computational overhead for real-time multi-vehicle control. To alleviate the above limitations, this paper proposes Enhanced Simulation environment based on Distribution matching and Large language model-Guided reinforcement learning (ESDLG), a two-stage simulation enhancement framework for background-vehicle policy learning. Rather than introducing a new divergence measure or a new reinforcement learning algorithm, ESDLG introduces χ^2-divergence-based state occupancy matching and uses LLM-guided policy learning for long-term traffic simulation enhancement. In the first stage, inspired by state occupancy matching formulations, ESDLG introduces a χ^2-divergence dual optimization module into PPO to provide a real-data-driven distribution-matching prior without requiring expert action labels. In the second stage, real trajectories are clustered into three distinct driving styles, and three style-specific PPO policies are fine-tuned using LLM-guided exploration. The LLM guidance is used only during training and distilled into lightweight policy networks; during deployment, the LLM is completely removed from the simulation loop. Experiments conducted in highway-env with HighD data show that ESDLG achieves balanced traffic-statistic fidelity, obtaining the best spacing MAPE and the second-best Hellinger distance for speed-distribution matching among the compared methods. ESDLG also preserves competitive coarse-grained driving-style diversity while reducing the per-step inference time to 15.9 ms, corresponding to a 5,424$\times$ speedup over online LLM calibration. Additional evaluations on traffic density and style ratio further indicate that ESDLG remains robust under varying traffic densities and supports soft scenario-level controllability through policy assignment.

1 INTRODUCTION

Autonomous driving simulation environments are central to algorithm development and safety validation, offering advantages in cost, scenario coverage, and risk controllability over real-road testing. However, during extended closed-loop operation, the statistical characteristics of simulated traffic flow—including speed distributions and inter-vehicle spacing—progressively diverge from real traffic data Wang et al. (2025), eroding the fidelity of simulation-based validation.

Existing enhancement methods fall into three categories. Rule-based models such as IDM + MOBIL form the default backbone of simulators like highway-env Leurent (2018); they are efficient but rely on hand-crafted parameters that cannot adapt to real-world distributions. Data-driven imitation learning methods Zhu et al. (2023); Xu et al. (2023) learn from real trajectories but suffer from compounding errors during closed-loop rollouts. LLM-based approaches Xu et al. (2025b); Pang et al. (2026) provide online corrective guidance but incur tens of seconds of inference per decision,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

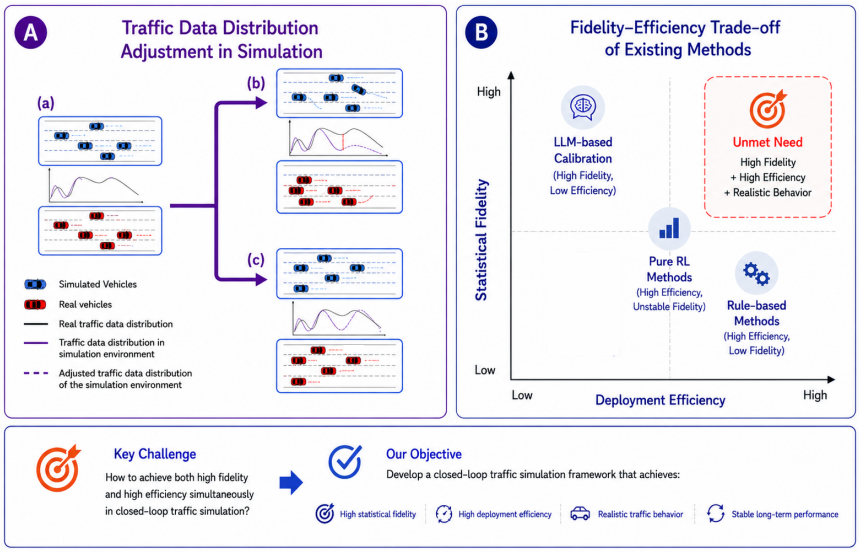

Figure 1: Motivation for improving the fidelity and computational efficiency of closed-loop background traffic simulation. Distribution correction is learned during training, so the simulator does not require an online LLM.

impractical for real-time multi-vehicle control where dozens of vehicles must be controlled each step.

These approaches face a fundamental fidelity–efficiency tension (Fig. 1): rule-based methods deploy instantly but lack distributional accuracy; LLM-based methods offer sophisticated correction at prohibitive cost; pure RL methods enable fast inference but produce policies whose state occupancy deviates from real data without explicit statistical constraints Ma et al. (2022).

We propose ESDLG, a two-stage framework following the route of "training RL policies $\rightarrow$ controlling background vehicles $\rightarrow$ enhancing simulation." Stage 1 embeds a χ^2-divergence dual optimization module (derived from SMODICE Ma et al. (2022)) into PPO: a dual value network minimizes divergence between policy-induced and real-traffic state occupancy, producing importance weights that reweight the PPO advantage—without requiring expert action labels. Stage 2 clusters real HighD trajectories into three driving styles and fine-tunes style-specific PPO policies via LLM-guided exploration with an adaptive exploration-imitation fusion loss Xu et al. (2025b). The LLM is used only during training and fully offloaded at deployment, where three lightweight policy networks collectively control background vehicles.

The main contributions are: (1) ESDLG integrates χ^2-divergence state occupancy matching into PPO and LLM-guided multi-style RL to mitigate statistical drift; (2) a multi-style training paradigm uses clustered real trajectories to train style-specific policies, enabling behavioral diversity without real-time LLM inference; (3) comprehensive experiments on highway-env with HighD data show ESDLG achieves the best spacing accuracy and competitive speed-distribution fidelity while substantially reducing deployment latency.

2 Related Work

Simulation environment enhancement. highway-env defaults to IDM + MOBIL, which fails to capture real traffic statistical regularities Krajewski et al. (2018). Data-driven methods (RITA Zhu et al. (2023), BITS Xu et al. (2023), HAD-Gen Wang et al. (2025), CtRL-Sim Rowe et al. (2024)) improve trajectory realism via imitation or offline RL but rely on expert demonstrations and do not address closed-loop distribution shift. Diffusion-based methods (SceneDiffuser Jiang et al. (2024), RIFT Chen et al. (2025)) focus on scene generation rather than sustained background vehicle control. We directly train RL policies with explicit macroscopic distribution constraints.

State occupancy matching. The DICE family Nachum et al. (2019) estimates stationary distribution corrections without expert actions; SMODICE Ma et al. (2022) formulates it as f-divergence matching via Fenchel duality, using χ^2-divergence with analytic importance weights. Extensions Yan et al. (2024); Yu et al. (2023); Sikchi et al. (2024); Seo et al. (2024) improve robustness but operate fully offline. We embed SMODICE's χ^2-divergence dual optimization into online PPO training via a V-dice network and importance-weighted advantages.

LLM-guided RL for driving. LEAD Xu et al. (2025b) uses dynamic LLM intervention with an AEIFL loss; TeLL-Drive Xu et al. (2025a), LGDRL Pang et al. (2026), VLM-RL Huang et al. (2025), and HCRMP Chen & Deng (2025) explore similar teacher–student paradigms, primarily for single ego-vehicle decisions. Online LLM inference remains costly for multi-vehicle simulation. We extend LLM guidance to multi-vehicle background control and fully offload the LLM via policy distillation.

Driving style clustering. Dynamic clustering de Zepeda et al. (2021) and subsequent work Yao et al. (2024); Lu et al. (2023); Li et al. (2025); Kushwaha et al. (2025) identify heterogeneous driving behaviors but are rarely integrated into simulation. We use clustered styles from real data to train concurrent style-specific RL policies.

3 METHOD

3.1 PROBLEM FORMULATION

We consider a multi-lane highway simulation with N background vehicles governed by policy $\pi(a|s)$. At step t, state $s_t \in \mathcal{S}$ contains kinematic states of observed vehicles over horizon T. The discounted state occupancy is $d^\pi(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr(s_t = s|\pi)$, and $d_E(s)$ is the empirical distribution from HighD. Under default IDM+MOBIL, $\hat{P}_{\text{sim}}(s)$ drifts from $P_{\text{real}}(s)$ in speed and spacing distributions. We learn K policies $\{\pi_k\}_{k=1}^K$ such that the aggregated occupancy $d_{\{\pi_k\}}(s)$ minimizes divergence to $d_E(s)$ while preserving diverse microscopic behaviors.

3.2 FRAMEWORK OVERVIEW

ESDLG uses two stages (Fig. 2). Stage 1 trains a single PPO policy with χ^2-divergence distribution matching to establish a macroscopic prior. Stage 2 replicates the checkpoint into $K = 3$ style-specific policies, fine-tuned with LLM guidance per driving-style cluster. Stage 1 regularizes toward real-traffic statistics; Stage 2 improves diversity and closed-loop consistency through style-guided supervision.

3.3 STAGE 1: χ^2-DIVERGENCE STATE OCCUPANCY MATCHING

The empirical speed distribution from HighD is discretized into $K_{\text{bin}} = 8$ equal-width buckets over $[5, 32]$ m/s:

$$p_k = \frac{\sum_{v\in\mathcal{V}} \mathbb{I}(v \in \mathcal{B}_k)}{\sum_{j=1}^{K_{\text{bin}}} n_j} \tag{1}$$

where $\mathcal{V}$ is the set of valid speed samples and $\mathcal{B}_k$ is the k-th interval. We choose χ^2-divergence for its convex quadratic dual loss and analytic importance weights, avoiding KL's logarithmic instability and Wasserstein's adversarial optimization loop.

Following SMODICE, a dual value network V_{dice} minimizes:

$$\begin{aligned}\mathcal{L}_{\text{dual}}(V) =& (1-\gamma)\mathbb{E}_{s\sim\mu_0}[V(s)] \\ &+ \mathbb{E}_{(s,a)\sim d^O}\left[\tfrac{1}{2}(R(s) + \gamma V(s') - V(s) + 1)^2\right]\end{aligned} \tag{2}$$

where $R(s) = \log\frac{d^E(s)}{d^O(s)}$ is reward shaping (set to $-C$ with $C = 10$ for out-of-distribution states), and μ_0 is the initial state distribution. The optimal importance weights are obtained analytically:

$$\xi^*(s,a) = \max(0,\ R(s) + \gamma V^*(s') - V^*(s) + 1) \tag{3}$$

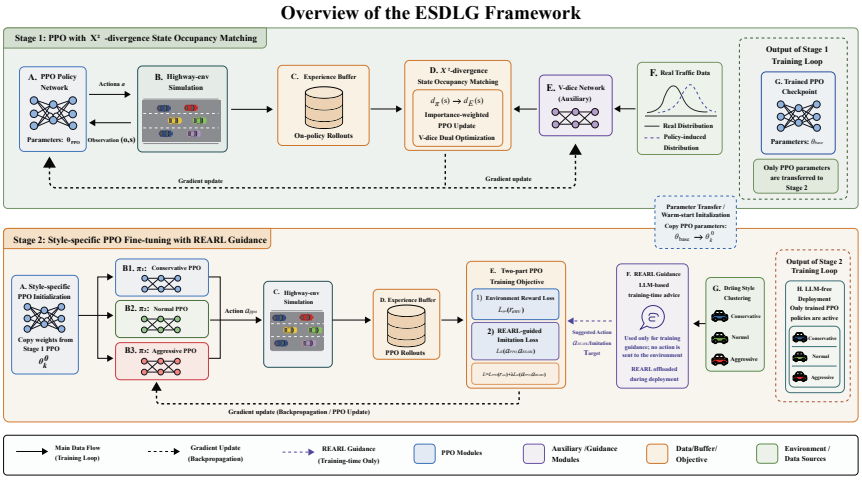

Figure 2: Overview of ESDLG. Stage 1 aligns policy-induced state occupancy with real traffic statistics through χ^2-divergence dual optimization. Stage 2 initializes style-specific PPO policies from the Stage 1 checkpoint and fine-tunes them with LLM supervision. During deployment, the LLM is offloaded and only trained PPO policies are used.

These weights reweight the PPO advantage:

$$\mathcal{L}_{\text{PPO}}(\theta) = \hat{\mathbb{E}}_t\big[\min\big(r_t(\theta)\xi^*\hat{A}_t, \\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\xi^*\hat{A}_t\big)\big] \tag{4}$$

where $r_t(\theta)$ is the policy probability ratio and $\hat{A}_t$ is the original advantage. This biases updates toward states matching real data without expert action labels. The base reward combines speed, collision, and lane-discipline terms:

$$r(s,a) = w_1 r_{\text{speed}} + w_2 r_{\text{collision}} + w_3 r_{\text{right_lane}} \tag{5}$$

with $w_1 = w_2 = w_3 = 1.0$; $r_{\text{collision}} = -1.0$ and $r_{\text{right_lane}} = +0.1$.

3.4 Driving Style Clustering from Real Trajectory Data

Between stages, we cluster HighD trajectories using an 8-dimensional feature vector:

$$\mathbf{f}_i = (\mu_v, \sigma_v, CV_v, \mu_s, \sigma_s, CV_s, \mu_{thw}, \sigma_{thw}) \tag{6}$$

comprising speed and spacing mean/std/coefficient of variation plus time headway mean/std. K-Means with $K \in \{3, 5\}$ and k-means++ initialization ($n_{\text{init}} = 20$) yields style clusters ordered by an aggressiveness score:

$$\text{score}_k = z(\sigma_v^{(k)}) + z(CV_v^{(k)}) - z(\mu_s^{(k)}) - z(\mu_{thw}^{(k)}) \tag{7}$$

where $z(\cdot)$ denotes z-score normalization. The 3-cluster model (Conservative/Moderate/Aggressive) provides training labels; the 5-cluster model serves as a fine-grained evaluator (Supplementary Section A). For each style, the $M = 10$ scenes closest to the centroid are selected as LLM training contexts.

3.5 Stage 2: Multi-Style LLM-Guided Dynamic Exploration

Twelve controlled vehicles are split into 3 groups of 4, each controlled by a style-specific policy π_{θ_k}. Every 5 simulated seconds, deviation detection computes the Hellinger distance between the current speed distribution P_t and the real distribution Q:

$$H(P_t, Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{K_{\text{bin}}}(\sqrt{p_{t,i}} - \sqrt{q_i})^2} \tag{8}$$

Table 1: Driving-style labels used in training and evaluation; the full feature descriptions are given in Supplementary Section A.

Cluster count	Use	Style labels
$K = 3$	Policy training	Conservative, Moderate, Aggressive
$K = 5$	Fine-grained evaluation	Very Conservative, Conservative, Moderate, Aggressive, Very Aggressive

along with spacing MAPE. When either exceeds its threshold, LLM intervention is triggered: the current scene plus representative style context is encoded and passed to the LLM, which outputs high-level actions (LANE_LEFT, SLOWER, IDLE, FASTER, LANE_RIGHT). LLM actions are distilled via cross-entropy imitation loss:

$$\mathcal{L}_{\mathrm{IL}} = \sum_{i \in \mathcal{M}} H(\delta(a_i^{\mathrm{LLM}}), \pi_\theta(\cdot \mid s_i)) \tag{9}$$

where $\mathcal{M}$ is the set of LLM-intervention samples. The total loss combines PPO and imitation objectives:

$$\mathcal{L}_{\mathrm{total}}(\theta) = \mathcal{L}_{\mathrm{PPO}}(\theta) + \omega \cdot \mathcal{L}_{\mathrm{IL}}, \quad \omega = 0.5 \tag{10}$$

The detailed training procedure is given in Supplementary Section B. At deployment, the LLM is removed; three PPO policies control all vehicles at $\sim$15.9 ms/step.

The three policies share the distribution-matched initialization but receive distinct supervision after style assignment. At each deviation check, the speed histogram is compared with the HighD target while spacing MAPE supplies a complementary gap-based signal. An intervention is recorded only if one of these two statistics crosses its threshold; the action source is stored alongside the transition, allowing the cross-entropy term to train on LLM-labelled states. Policy actions are used directly when the thresholds are not exceeded. This training arrangement separates corrective supervision from deployment: the simulator consults only policy networks after fine-tuning, including in the long-horizon evaluation.

3.6 Complexity Analysis

ESDLG deployment complexity is $\mathcal{O}(TN_cC_\pi)$ with memory $\mathcal{O}(K|\theta_\pi|)$, independent of LLM latency. Online LLM calibration adds $\mathcal{O}(\rho TMC_{\mathrm{LLM}})$, which dominates since $C_{\mathrm{LLM}} \gg C_\pi$. ESDLG shifts this cost to training, where LLM guidance is distilled into policies and disappears after distillation.

4 Experiments

4.1 Experimental Setup

Platform. highway-env Leurent (2018), four-lane highway, 50 initial vehicles, density 1.0, 60 s duration at 15 Hz ($\Delta t = 0.0667$ s). Kinematics observation (5 vehicles, normalized relative coordinates); DiscreteMetaAction (9 target speeds in $[5, 32]$ m/s). Background vehicles not under RL control follow default IDM+MOBIL.

Data. HighD Krajewski et al. (2018) (16.5 h, $\sim$110,000 vehicles, 5,600 lane changes); speed samples discretized into 8 buckets over $[5, 32]$ m/s. Two K-Means models ($K = 3, 5$) are trained on trajectory features; the 3-cluster model provides training labels, the 5-cluster model is retained as a fine-grained evaluator.

Training. Stage 1: 200,000 steps, lr 5×10^{-4}, update every 256 steps (10 epochs, batch 10). Stage 2: 10,000 steps per style, update every 60 steps (batch 32, 50 epochs); deviation detection every 5 decision frames (1 frame = 15 simulation frames = 1 s). 5 random seeds, averaged over 50 test scenarios.

Metrics. Speed Hellinger distance (HD, lower), spacing MAPE (lower), per-step inference time, mode entropy H_{mode} (higher, style diversity), and Jensen–Shannon distance to real style distribution JS_{real} (lower). Style metrics computed under both 3-cluster and 5-cluster evaluators.

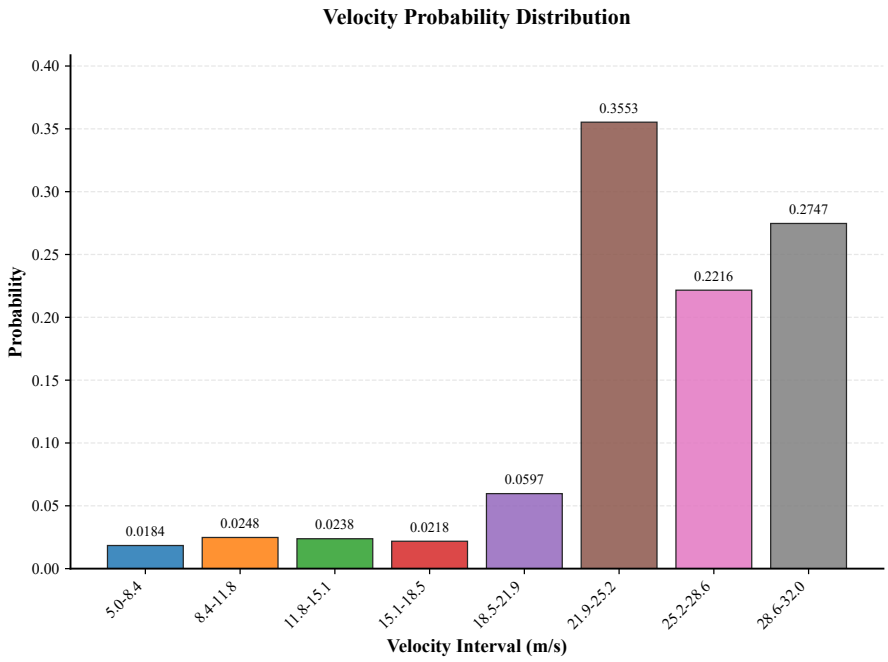

Figure 3: HighD speed histogram defining the eight-bin empirical reference used in Stage 1.

Table 2: Comparison of divergence and driving-style metrics across methods. Best bolded, second-best underlined.

Category	Setting	Metric	REARL	PPO	DICE	ESDLG
Divergence	Traffic stats	Speed (HD) ↓	**0.3067**	0.3258	0.3470	<u>0.3226</u>
		Spacing (MAPE) ↓	<u>0.8371</u>	1.3940	0.8812	**0.8089**
Style metric	3-cluster	H_{mode} ↑	<u>0.8742</u>	0.6932	0.7855	**0.8748**
		JS_{real} ↓	**0.2092**	0.3198	0.2808	<u>0.2171</u>
		Conservative	0.5666	0.7350	0.6466	0.5875
		Moderate	0.1450	0.1383	0.2500	0.2291
		Aggressive	0.2883	0.1266	0.1033	0.1833
	5-cluster	H_{mode} ↑	**0.8559**	0.8071	<u>0.8339</u>	0.7826
		JS_{real} ↓	**0.3916**	<u>0.4028</u>	0.4285	0.4848
		Very Conservative	0.3050	0.3600	0.4516	0.4875
		Conservative	0.2883	0.4066	0.2283	0.1055
		Moderate	0.1366	0.1100	0.2100	0.2180
		Aggressive	0.0066	0.0266	0.0483	0.0055
		Very Aggressive	0.2633	0.0966	0.0616	0.1833

Baselines. Base (IDM+MOBIL), REARL (online LLM calibration), PPO Schulman et al. (2017) (standard RL without distribution matching or LLM), DICE (Stage 1 only, χ^2-divergence matching without LLM guidance).

4.2 Performance Comparison

Table 2 and Fig. 4 compare all methods. ESDLG achieves the best spacing MAPE (0.8089, outperforming REARL by 3.37%, DICE by 8.20%, PPO by 41.97%) and second-best speed HD (0.3226 vs. REARL's best 0.3067). The χ^2-divergence module alone (DICE) yields higher speed HD (0.3470), confirming it serves as a regularization prior whose benefit emerges with Stage 2 refinement.

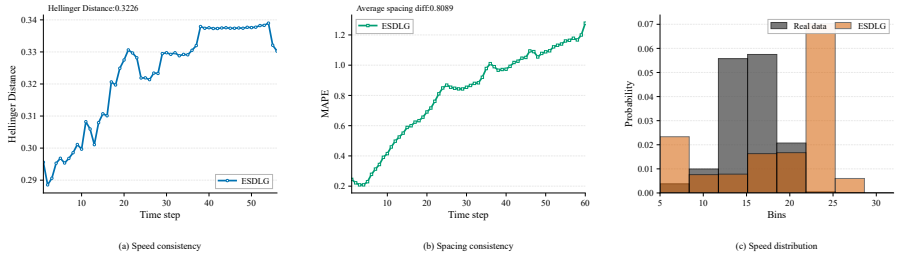

Figure 4: Macroscopic traffic-statistic fidelity. (a) Temporal evolution of speed Hellinger distance. (b) Temporal evolution of spacing MAPE. (c) Speed-bin distribution comparison between HighD and ESDLG.

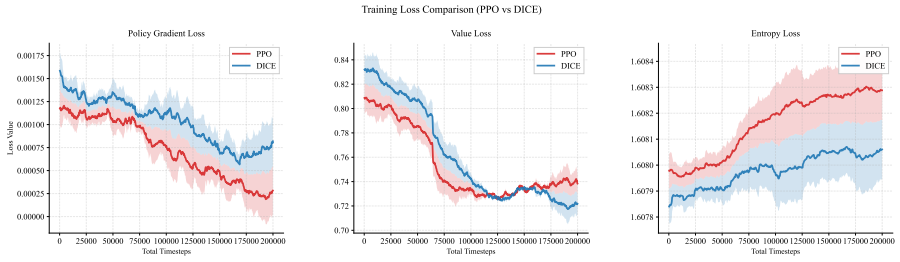

Figure 5: Training loss comparison between standard PPO and PPO with χ^2-divergence state occupancy matching over 200,000 steps.

Under the 3-cluster evaluator, ESDLG obtains the highest H_{mode} (0.8748, comparable to REARL's 0.8742) and second-best JS_{real} (0.2171), reducing PPO's excessive conservative concentration (0.7350) to 0.5875 and increasing Moderate/Aggressive proportions. Under the 5-cluster evaluator, REARL remains stronger; ESDLG's three-policy design underrepresents fine-grained styles (48.75% Very Conservative). Overall, ESDLG is an efficiency-oriented trade-off: it preserves coarse-grained fidelity and diversity without online LLM inference.

What each baseline isolates. The standard PPO baseline evaluates policy learning without occupancy correction or LLM feedback. DICE retains the first stage and therefore isolates the effect of adding the distribution-matching prior before style-specific fine-tuning. ESDLG adds the second stage, while REARL provides the online-LLM comparison for both fidelity and inference cost. The speed HD values for PPO, DICE and ESDLG are 0.3258, 0.3470 and 0.3226, respectively. DICE alone therefore does not improve the speed histogram; the final result is attributable to the two-stage configuration in this comparison rather than to a monotonic improvement from each individual stage. For spacing MAPE the same methods score 1.3940, 0.8812 and 0.8089, so the distribution-matched initialization supplies a clearer benefit in spacing before LLM-guided refinement. REARL achieves the lowest speed HD, 0.3067, but requires 86,241.309 ms per step, compared with 15.898 ms for ESDLG. The comparison thus identifies a concrete fidelity and latency trade-off rather than a uniform win across every metric.

4.3 Training Stability Analysis

Fig. 5 compares PPO and DICE training losses over 200,000 steps. Both converge stably without divergence. DICE shows slightly higher mid-training variance in policy gradient loss but comparable value loss convergence ($\sim$0.72 vs. $\sim$0.74 for PPO), confirming that χ^2-divergence regularization does not destabilize PPO. Entropy dynamics remain well controlled for both methods.

4.4 Inference Time and Computational Complexity

Table 3 shows REARL requires 86,241.309 ms/step with online LLM, while ESDLG needs only 15.898 ms—a 5,424$\times$ speedup. Base (rule-based) is fastest at 0.461 ms. ESDLG converts LLM-

Table 3: Per-step inference time comparison.

Method	Inference (ms)	Speedup vs. REARL
Base	0.461	—
REARL (LLM online)	86241.309	1×
ESDLG (Ours)	**15.898**	**5424**×

guided enhancement from an ∼86-second online process to millisecond-level policy inference, suitable for real-time multi-vehicle simulation (highway-env requires 66.7 ms/step at 15 Hz).

4.5 Sensitivity to Traffic Density

We evaluate at densities 0.5 (sparse), 1.0 (default), and 2.0 (dense) (Supplementary Section D). ESDLG is most stable at density 1.0, achieving the best spacing MAPE (0.8089) and lowest 3-cluster JS_{real} (0.2171). Denser traffic increases speed HD (0.3428) and spacing MAPE (1.1374) due to stronger interactions, but also raises behavioral diversity ($H_{\text{mode}} = 0.9579$ at 3-cluster). Sparse traffic has slightly lower speed HD (0.3184) but higher spacing error (0.9562).

Table 4: Key traffic statistics across evaluation densities. The complete style breakdown is in Supplementary Section D.

Density	0.5	1.0	2.0
Speed HD ↓	0.3184	0.3226	0.3428
Spacing MAPE ↓	0.9562	0.8089	1.1374
3-style entropy ↑	0.8232	0.8748	0.9579
3-style JS ↓	0.2586	0.2171	0.2465

4.6 Style-Ratio-Based Scenario Evaluation

We construct five scenarios by assigning Conservative/Moderate/Aggressive policies at different ratios (Supplementary Section E), testing whether learned policies enable scenario-level controllability. Each scenario runs 5 episodes (60 decision frames each) with deterministic inference.

Results (Supplementary Section E) show speed HD remains stable across scenarios (0.3191–0.3319), while spacing MAPE varies more (0.7315–1.4287). Under 3-cluster evaluation, the Aggressive scenario achieves the highest diversity ($H_{\text{mode}} = 0.9369$) and closest real-style match ($JS_{\text{real}} = 0.1578$); the Very Aggressive scenario produces the largest Aggressive proportion (0.3403). Under 5-cluster, the Very Aggressive scenario best matches fine-grained styles ($H_{\text{mode}} = 0.8504$, $JS_{\text{real}} = 0.3777$). Different ratios produce distinguishable distributions, demonstrating soft scenario-level controllability, though final behavior is also shaped by policy dynamics and traffic interactions.

Table 5: Selected measures for five policy-assignment scenarios. Complete distributions and assignment counts are given in Supplementary Section E.

Scenario	V. cons.	Cons.	Moderate	Aggressive	V. aggr.
Speed HD	0.3276	0.3191	0.3224	0.3284	0.3319
Spacing MAPE	0.8938	0.9745	0.7315	1.1310	1.4287
3-style JS	0.2160	0.2103	0.2331	0.1578	0.1806

Interpreting the stress tests. The density sweep tests whether the distribution-matched initialization continues to help when vehicle interactions become more frequent. At density 2.0, both speed HD and spacing MAPE worsen relative to density 1.0, so the observed gain at the default density should not be interpreted as density-invariant fidelity. Style entropy instead rises with density, indicating that this diversity statistic does not by itself imply better agreement with the HighD reference;

the JS measure and traffic statistics are required to interpret it. The policy-ratio experiment probes a distinct question: changing the policy mix changes the resulting style distribution, but it does not prescribe an exact output distribution. Its spacing error also varies more than its speed HD, which limits the precision of scenario-level control. These stress tests therefore complement the default-density comparison rather than replacing it.

4.7 Discussion of Fidelity and Deployment

The evaluation combines aggregate traffic statistics with style-sensitive measures because these capture different aspects of the background fleet. Speed HD compares histogram shape, whereas spacing MAPE responds to errors in inter-vehicle gaps. Neither guarantees that a policy maintains all microscopic properties of an individual HighD trajectory. The higher 3-cluster mode entropy of ESDLG indicates a more dispersed coarse style mixture than PPO; the 5-cluster JS result is less favorable and exposes a limitation of representing the data with only three learned policies. These observations constrain the interpretation of "behavioral diversity": it refers to the reported cluster-level outcomes, not a comprehensive guarantee of realistic interactions.

Deployment timing addresses a different bottleneck. ESDLG removes LLM requests after training and runs at 15.898 ms per simulation step in the reported configuration. This is below the 66.7 ms interval corresponding to 15 Hz, whereas online REARL calibration takes 86,241.309 ms per step. The measurements support inference efficiency for this specific simulator and vehicle count; they do not establish identical latency for larger fleets, different hardware or more complex road layouts. In addition, the experiments use highway-env and HighD, so transfer to urban intersections, rare maneuvers and other trajectory collections remains untested.

The density and policy-ratio studies help delimit generalization within the evaluated highway setting. At density 2.0, spacing MAPE reaches 1.1374, above the 0.8089 default-density value. Mixture assignments vary the realized style distribution but also affect spacing error. Hence an operator should treat style assignment as a soft scenario control input and assess traffic statistics after rollout rather than infer exact behavior from assignment counts alone. These are empirical observations from the present evaluation and motivate more adaptive style partitions and direct closed-loop constraints in future work.

5 Conclusion and Future Work

ESDLG enhances autonomous driving simulation through χ^2-divergence distribution matching and LLM-guided multi-style RL, achieving the best spacing fidelity and competitive speed-distribution matching while reducing deployment inference to 15.9 ms (5,424$\times$ over online LLM calibration). Density and style-ratio experiments confirm robustness under varying traffic conditions and soft scenario-level controllability. Limitations include restriction to highway scenarios and insufficient fine-grained style recovery with only three policies. Future work will extend to urban intersections and roundabouts, investigate adaptive style partitioning and more expressive multi-policy architectures, and validate on additional real-world trajectory datasets.

AI Use Statement

The authors used generative AI tools to assist with language polishing and conversion to LaTeX format. All scholarly content of this paper, including the research questions, research methods, experimental procedures, numerical results, and conclusions, was independently determined and verified by the authors. The authors have reviewed all AI-assisted content and take full responsibility for the final submitted manuscript.

REFERENCES

Chen, Sun, Cheng, and Zheng. Closed-loop rl fine-tuning for realistic and controllable traffic simulation. arXiv preprint arXiv:2505.03344, 2025.

Zhiwen Chen and Hanming Deng. Hcrmp: An llm-hinted contextual reinforcement learning framework for autonomous driving. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Maria Valentina Nino de Zepeda, Fanlin Meng, Jinya Su, Xiao-Jun Zeng, and Qian Wang. Dynamic clustering analysis for driving styles identification. *Engineering Applications of Artificial Intelligence*, 97:104096, 2021. doi: 10.1016/j.engappai.2020.104096.

Zilin Huang, Zihao Sheng, and Yansong Qu. A unified vision language models and reinforcement learning framework for safe autonomous driving. *Transportation Research Part C: Emerging Technologies*, 2025. doi: 10.1016/j.trc.2025.105032.

Max Jiang, Yijing Bai, and Andre Cornman. Scenediffuser: Efficient and controllable driving simulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems. In *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2018. doi: 10.1109/ITSC.2018.8569552.

A. Kushwaha, P. Kumar, and S. Arora. A deep embedded clustering method for location-specific driving safety profiling using trajectory data. *Scientific Reports*, 15, 2025. doi: 10.1038/s41598-025-25538-8.

Edouard Leurent. An environment for autonomous driving decision-making. GitHub repository, 2018. https://github.com/Farama-Foundation/HighwayEnv.

Bingxian Li, Yanfang Liu, Junwei Zhao, Xiangyang Xu, and Peng Dong. Representing-behavior-based online driving style recognition and its road verification. *Engineering Applications of Artificial Intelligence*, 2025. doi: 10.1016/j.engappai.2025.012424.

Lin Lu, Yao Lin, Yuan Wen, Jinxiong Zhu, and Shengwu Xiong. Federated clustering for recognizing driving styles from private trajectories. *Engineering Applications of Artificial Intelligence*, 118:105714, 2023. doi: 10.1016/j.engappai.2022.105714.

Yecheng Jason Ma, Andrew Shen, Dinesh Jayaraman, and Osbert Bastani. Versatile offline imitation from observations and examples via regularized state-occupancy matching. In *International Conference on Machine Learning (ICML)*, pp. 14639–14663. PMLR, 2022. doi: 10.48550/arXiv.2202.02433.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Hao Pang, Zhenpo Wang, and Guoqiang Li. Large language model guided deep reinforcement learning for safe autonomous vehicle decision making. *Transportation Research Part C: Emerging Technologies*, 184:105511, 2026. doi: 10.1016/j.trc.2025.105511.

Luke Rowe, Roger Girgis, and Anthony Gosselin. Ctrl-sim: Reactive and controllable driving agents with offline reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Seokin Seo, Byung-Jun Lee, Jongmin Lee, Hongseok Yang, and Kee-Eung Kim. Mitigating covariate shift in behavioral cloning via robust stationary distribution correction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Harshit Sikchi, Caleb Chuck, Amy Zhang, and Scott Niekum. A dual approach to imitation learning from observations with offline datasets. In *Conference on Robot Learning (CoRL)*, 2024.

Cheng Wang, Lingxin Kong, Massimiliano Tamborski, and Stefano V. Albrecht. Human-like and diverse driving behavior modeling for controllable scenario generation. *Accident Analysis & Prevention*, 223:108270, 2025. doi: 10.1016/j.aap.2025.108270.

Chengkai Xu, Jiaqi Liu, Shiyu Fang, Yiming Cui, Dong Chen, Peng Hang, and Jian Sun. Tell-drive: Enhancing autonomous driving with teacher llm-guided deep reinforcement learning. *arXiv preprint arXiv:2502.01387*, 2025a.

Danfei Xu, Yuxiao Chen, Boris Ivanovic, and Marco Pavone. Bi-level imitation for traffic simulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2929–2936, 2023.

Dongwei Xu, Enwen Qiao, Tongcheng Gu, Hongda Fu, Chengju Sun, Haifeng Guo, and Yuqing Liu. Lead: Llm-enhanced deep reinforcement learning for stable decision-making in critical autonomous driving scenarios. *Neurocomputing*, 657, 2025b. doi: 10.1016/j.neucom.2025.129209.

Kai Yan, Alexander G. Schwing, and Yu-Xiong Wang. Offline imitation from observation via primal wasserstein state occupancy matching. In *International Conference on Machine Learning (ICML)*. PMLR, 2024.

Yiru Yao, Simeon Calvert, and Serge Hoogendoorn. Driving pattern interpretation based on action phases clustering. *arXiv preprint arXiv:2407.17518*, 2024.

Lantao Yu, Tianhe Yu, Jiaming Song, Willie Neiswanger, and Stefano Ermon. Offline imitation learning with suboptimal demonstrations via relaxed distribution matching. In *AAAI Conference on Artificial Intelligence*, 2023. doi: 10.1609/aaai.v37i9.26305.

Zhengbang Zhu, Shenyu Zhang, and Yuzheng Zhuang. Boost autonomous driving simulators with realistic interactive traffic flow. In *Distributed Artificial Intelligence (DAI)*, 2023.