Diverse Human Driving Vehicle Simulation in Background Traffic for Autonomous Driving Tests

Wendi Li[1], Hao Wu[1], Han Gao[1], Bing Mao[1], Fengyuan Xu[1]*, Sheng Zhong[1]

[1]National Key Lab for Novel Software Technology, Nanjing University
wendili@smail.nju.edu.cn, hao.wu@nju.edu.cn, gaohan@smail.nju.edu.cn, maobing@nju.edu.cn, fengyuan.xu@nju.edu.cn, zhongsheng@nju.edu.cn

Abstract

Realistic background traffic is critical to the simulation platforms for autonomous driving (AD) testing. Given that most vehicles in reality are driven by human beings, introducing human driving (HD) vehicles to the background traffic is necessary to be able to discover more problems of the tested AD vehicle in the simulation stage. However, existing methods rely on ad-hoc rules or data-driven training to mimic partial human driver behaviors, which are not comprehensive and lack transparency. In this work, we design a smart human driving vehicle simulator `HDSim` which is empowered by cognitively inspired modeling and AI models. `HDSim` enables diverse, realistic, and scalable HD traffic simulation on AD testing platforms like CARLA in a non-intrusive manner. There are two novel components in `HDSim`. First, we introduce a driver model to guide the generation of diverse human driving styles by using different combinations of latent cognitive factors in a hierarchy. Second, we design a Perception-Mediated Behavior Influence (PMBI) mechanism to use LLM-assisted perceptual transformations to indirectly fuse driving actions with driving styles. Experiments show that `HDSim` traffic can help simulation platforms like CARLA to reveal 68% more failures of tested AD vehicles, and the explainability of reported accidents is also improved.

Introduction

Simulation constitutes a critical testing phase prior to the real-world deployment of any end-to-end autonomous-driving (AD) vehicle. Enhancing the realism of background traffic is a core task for existing simulation platforms such as CARLA. Among all traffic characteristics, the inclusion of human-driving (HD) vehicles is particularly significant, as it introduces a unique source of behavioral variability. Compared to simplified traffic scenarios, traffic mixed with human-driving vehicles can expose more issues in AD vehicles under test. Figure 1 demonstrates the key advantage of such mixed traffic and presents an example identified by using our simulator `HDSim`.

However, existing background vehicle simulations fall short in modeling human drivers. Most rely on rule-based or replay-based methods (Behrisch et al. 2011; Dosovitskiy et al. 2017; Montali et al. 2023; Li et al. 2022), so they only support homogeneous HD behaviors and cannot cover diverse HD styles like aggression and driving under the influence (DUI). Recent approaches attempt to increase behavioral diversity through data-driven learning approaches like (Tan et al. 2024). Yet, the need for real-world training data results in limited support for diverse HD styles, since collecting data on abnormal behaviors is dangerous for real drivers. Besides, it remains unclear whether a driving model trained using stylized data, e.g., actions from aggressive drivers, could exhibit unexpected or illegal behaviors.

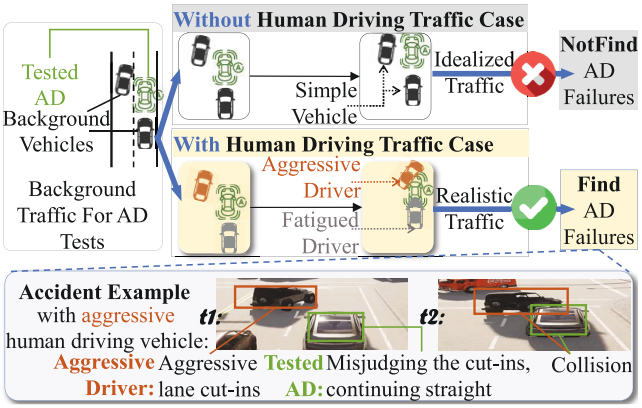


Figure 1: Diverse stylized traffic simulation reveals greater potential in identifying the shortcomings of AD models. Green box: tested AD, blue-white box: aggressive driver.

Therefore, in this work, we aim to explore the feasibility of introducing diverse HD vehicles into background traffic simulations. First, we expect these HD vehicles to operate as if driven by humans with competent driving skills and to interact with the AD vehicle under test using appropriate judgment. Moreover, we anticipate each of these vehicles to mimic assigned styles, such as driving and DUI, while ensuring behavioral explainability, which facilitates root cause identification later. Finally, this approach eliminates the need to collect real-world training data and can be easily integrated into existing simulation platforms like CARLA.

To achieve the above goals, we propose `HDSim`, a cognitively inspired framework for simulating interactive traffic with diverse human driving styles. An HD vehicle in `HDSim`is built upon an autonomous driving model, enabling

*Corresponding author.

it to interact with the vehicle under test more naturally than rule-based solutions. Beyond this, our core design comprises two key components. First, we propose a novel human driving style modeling approach to serve as a theoretical guideline. Second, we leverage a large language model (LLM) to non-intrusively adjust the perceptual inputs received by this autonomous driving model, in accordance with our proposed modeling. Such reconstruction of perceptual inputs (e.g., object size and speed) will modulate the driving behaviors shaped by the target style. Given that the LLM is pre-trained with essential knowledge in transportation and cognitive science, it can understand and adhere to our proposed modeling without the need for fine-tuning. More specifically, we address the following three challenges.

First, model latent human driving styles. We develop a hierarchical human driving style model rooted in cognitive science. This model disentangles basic driving capabilities from three composable style influence layers: personality, physiological state, and attentional response, each associated with distinct temporal patterns, enabling semantically rich style representations and modular composition of traits.

Second, inject perceptual biases via LLM-guided policy transformation. We propose a novel Perception-Mediated Behavior Influence (PMBI) mechanism. This mechanism influences driving behavior by altering the vehicle agent's perception through predefined API functions and LLM-based object-level adjustments regulated by threshold constraints, thereby modeling how cognitive traits bias attention and salience while keeping the AD model's logic unchanged.

Third, enable scalable integration and stable simulation. Our design is non-intrusive, requiring no architectural changes to AD models during diverse agents' simulation. It ensures stability by recording all activities for replay without LLMs, delegating rendering to the platform while using LLMs only to adjust object attributes such as location.

Experiments show that `HDSim` exposes substantial weaknesses in state-of-the-art AD models, identifying up to 68% more failures than traditional tests and improving the real-world explainability of reported accidents. It also generates stylistically diverse yet realistic traffic while supporting stable and scalable simulation.

Our contributions are summarized as follows:

- We are the first to model diverse stylized human driving behaviors within interactive traffic environments, significantly enhancing realism over existing rule-based or homogeneous simulation agents for AD evaluation.
- We propose a hierarchical human driver style model that disentangles style influences from basic driving capabilities, and implement it via a heterogeneous AI approach to simulate human-like driving behaviors.
- We design a simulation mechanism that is controllable, scalable, and computationally efficient, facilitating easy integration with existing AD simulators and straightforward adaptation to new driving styles without retraining.
- Extensive experiments on a state-of-the-art simulator show that incorporating stylized human drivers uncovers up to 68% more hidden failures in AD models than conventional testing, improving the real-world explainability of accidents, without any physical deployment.

`HDSim` Design

This section first introduces our cognitively-inspired human driving style model, and then describes how to realize such a model into a human driver simulator, which can be non-intrusively applied onto existing AD testing simulations like CARLA (Dosovitskiy et al. 2017), as illustrated in Figure 2.

Human Driving Style Model

We propose a hierarchical model supported by knowledge from cognitive behavioral science. This model extracts cognitive factors, which lead to subtle driving action diversity, from basic driving capabilities and organizes them into a hierarchy. Such a hierarchical structure can accumulate the effects of multiple cognitive factors and guide LLMs to diversify abstracted human driving styles without knowing the concrete driving actions beforehand.

Inspired by psychological theories of individual variability (Elander, West, and French 1993; Taubman-Ben-Ari, Mikulincer, and Gillath 2004a), the model is structured as a set of concentric layers (Figure 3), with an inner Driving Capability Layer (DCL) and three outer Style Influence Layers (SILs). This design supports modular composition of human-like driving behaviors in a scalable and semantically coherent way.

At the center of this model, the DCL represents the driving action decisions made by human drivers according to their perceived in-situ contexts on the road, such as stopping at a red light, avoiding obstacles, and turning at a crossing. We assume this is shared by all rational human drivers with driver permits. Apparently, this part can be easily realized by current mainstream AD models.

SILs, surrounding the DCL, encode those stackable driving effects of cognitive factors into three concentric rings, namely the personality influence layer, the physiological influence layer, and the attentional influence layer. The further a layer is from the center, the more transient and unstable its influence on the driving style is.

- L1 – Personality Influence Layer captures enduring psychological attributes, such as aggressiveness or cautiousness, that shape baseline risk perception and planning tendencies over long timescales. (Taubman-Ben-Ari, Mikulincer, and Gillath 2004b).
- L2 – Physiological Influence Layer models physiological states (e.g., fatigue, intoxication) that modulate behavioral control. It captures two forms of influence: *(i)* Incremental effects like fatigue that accumulate and alter behavior once a cognitive threshold is exceeded; and *(ii)* Episodic effects like alcohol impairment that disrupt judgment and reactivity (Philip et al. 2001).
- L3 – Attentional Influence Layer accounts for attention dynamics in response to external complexity, grounded in Attention Resource Theory (Kahneman 1973). Distracted driving, for instance, manifests as lagged awareness and missed hazards (Endsley 2017; Schwarz and Weller 2023).

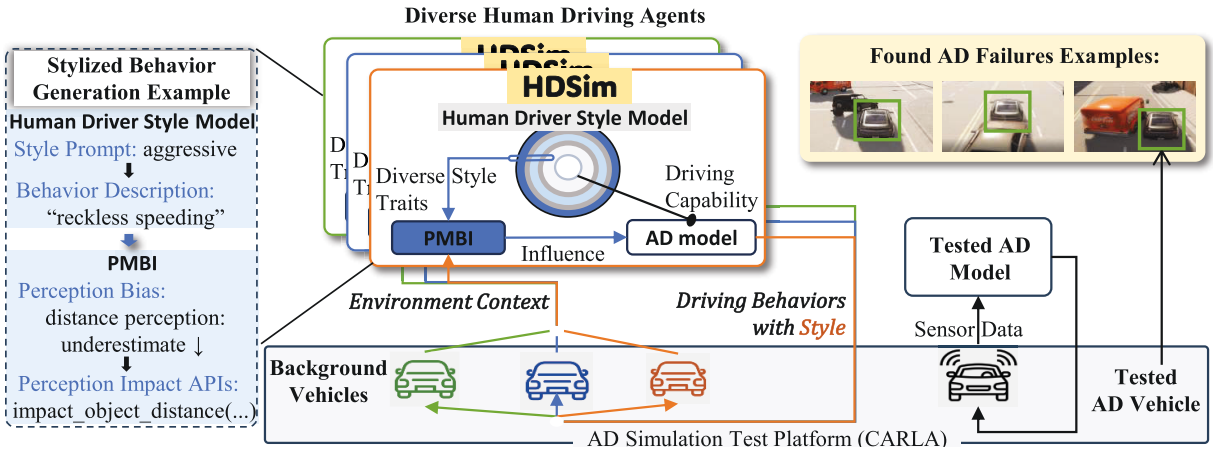


Figure 2: Overview of HDSim: a human driver simulation framework for generating background traffic populated by driver agents with diverse human-like driving styles, each making decisions based on stylized subjective perception.

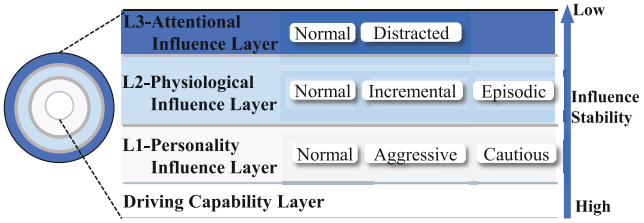


Figure 3: Hierarchical human driver style model.

Cognitive factors are categorized into these three layers based on their temporal stability (Sheeran, Orbell, and Trafimow 1999) and intensity (Norman and Shallice 1986), impacting the subtle driving action changes. A factor in the personality influence layer yields persistent driving characteristic changes, a factor in the physiological influence layer can be realized by periodical updates with a stylistic coherence mechanism, and a factor in the last layer captures unpredictable behavioral fluctuations, which can be realized as a stochastic process. When all factors in three layers are jointly modeled and coherently applied to concrete driving actions, the driving style is created for a simulated human driving vehicle.

Thanks to this modeling, we can assign different human driving styles to each simulated human driving vehicle without the association of concrete driving actions made by the vehicle. Such a human driving style assignment can be in the form of a semantic description, which is an expertise of LLMs. Therefore, we let an LLM learn our human driving style model via in-context learning and generate a targeted driving style description at a high level for every simulated human driving vehicle. This description ("An aggressive driver who accelerates through sparse traffic, tailgates consistently, and disregards speed limits with reckless confidence.") will further be used to guide the micromanipulation of every action of the assigned vehicle, achieving the targeted driving style influence, such as accelerating on a yellow light, hitting the brakes when the obstacle is far away, and turning by crossing solid lane marks.

Please note that, in this stage, the LLM is only used once to generate the driving style description, which is used in a plug-and-play manner to indirectly influence the driving (details are provided in the next subsection). No training is required, and the application is transparent to the AD testing simulations, like CARLA. Thus, the proposed model makes our human driver simulator design scalable and non-intrusive to existing simulation platforms.

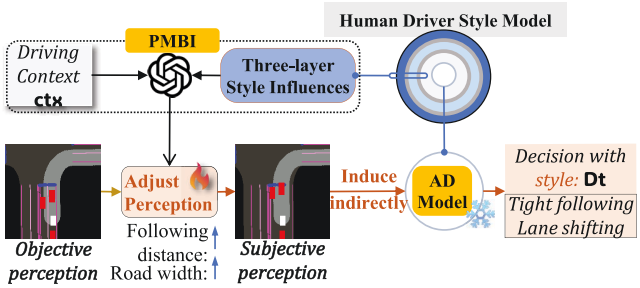


Figure 4: Stylized Driver Behavior Generation.

Stylized Driver Behavior Generation

For clarity, we first define the I/O of our stylized driver behavior generation. One input is a text description of the targeted human driving style $\mathcal{B}_{desc}$, which is provided by an LLM according to our human driving style model. $\mathcal{B}_{desc}$ is fixed during the simulation of style-assigned background vehicle. The other input is the context information $\mathbf{ctx}$ at the current moment (i.e., the current step), which is provided by the AD testing simulations. $\mathbf{ctx}$ contains all objects surrounding the simulated human driving vehicle (please note it is not the tested AD vehicle) and their physical states, like speeds and directions. On the other hand, the output of this stylized driver behavior generation is the driving action decision for the next step, which is denoted as $\mathbf{D_t}$.

Our HDSim design has two key components, as shown in

Figure 4. The first is an AD model taking as input the BEV (bird's eye view) scenes centered on it. HDSim leverages this AD model to realize the DCL in our human driving style model. It supports our HDSim vehicle to run from one place to another. Its driving actions, without the help of the second component, are denoted $\mathbf{D^{dcl}}$. The second component is the PMBI mechanism powered by an LLM, which realizes the remaining three layers in our driving style model. PMBI applies a targeted human driving style onto each $\mathbf{D_t^{dcl}}$ indirectly by manipulating the BEV scene perceived by the first component, following the view that behavioral differences arise from perception rather than ability bias (Elander, West, and French 1993). This non-intrusive approach preserves the correctness of the original AD model $\mathcal{M}_{AD}$ via the PMBI mechanism $\mathcal{M}_{AD}(\mathbf{ctx}')$ than direct policy modulation $\mathcal{M}'_{AD}(\mathbf{ctx})$, and offers clear explainability for human inspectors when accidents occur.

Together, they can generate a sequence of driving actions $\mathbf{D}$ with certain human driving characteristics, such as aggression, driving under the influence (DUI), and distraction. Additionally, HDSim-controlled vehicles can interact with the tested AD vehicle on the road, making the simulation more realistic. Existing AD testing simulation platforms can easily replace rule-based or non-responsive background vehicles with HDSim ones, without modifying the platforms.

Perception Mediated Behavior Influence (PMBI) Mechanism

In this subsection, we elaborate on our PMBI mechanism, which is critical in HDSim to naturally fuse the driving decision with the desired human driver style. Our design intuition has two points. First, we ask the AD model of our first component to take the lead in driving decisions, avoiding decision conflicts if there are multiple decision makers in the system. Second, we indirectly influence decisions of the AD model through the manipulation of what the model takes as input. In other words, we create an illusion gap between what is objective in the simulated context and what is subjective in the model's perception. This illusion gap will guide the AD model in adjusting its action decision a bit, which reflects the human driving style assigned to it.

The creation of this illusion gap is automated by an LLM with expert knowledge from cognitive behavioral science. We introduce a few principles and use them in the in-context learning for LLM. These principles teach the LLM how to interpret $\mathcal{B}_{desc}$ into a policy set $\mathcal{P} = (p_1, p_2, p_3)$, where p_1, p_2, and p_3 instruct in text how the personality, physiological, and attentional factors make subtle behavior changes of driving actions, respectively. For example, an aggressive personality characteristic like "confidently underestimates surrounding risks" is interpreted into a p_1 policy of "perceived distance of objects like vehicles in front is further than the real distance". For another example, a physiological characteristic like "DUI" is interpreted into a p_2 policy of "perceive straight lane marks as curved".

Since our AD model perceives context information in the BEV form, we provide a set of APIs with documentation to the LLM. These APIs are function calls designed to modify the contents of a BEV scene, such as changing object size, changing object location, and changing traffic lights, with sensitivity coefficients across motion, spatial, structural, and temporal dimensions (Green and Petre 1996). We also provide a RAG containing typical code examples of using these APIs to implement a policy p. Examples are fed into the in-context learning way to LLM so that LLM can write high-quality code for the desired policy. Thus, we give the generic LLM the ability to translate text-based policies into API-based code instructions.

During the simulation of a simulated human driving vehicle, translation is performed once for p_1 policies, periodically repeated with new configurations for p_2 policies, and randomly triggered for p_3 policies. Temporal consistency is preserved for all action decisions between two same-layer policy translations by verifying that linear changes remain physically plausible with respect to factors such as speed and size, and this procedure can be extended with other advanced physical models. Additionally, the LLM cost is saved.

Procedure Description. At the moment of making the t-th step driving action for one simulated vehicle, the procedure of PMBI is illustrated as follows. First, PMBI collects ctx_t information from the AD testing simulation platform, like CARLA. For every object shown in the defined BEV area of the simulated vehicle, PMBI matches suitable API-based code instructions to change how it looks in the eye of our AD model. The consistency of change is also considered in our API implementations. After that, the changed BEV is passed to the AD model of this vehicle to generate D_t. The procedure of overall stylized driver behavior generation with PMBI as the Algorithm 1.

Algorithm 1: Stylized Driver Behavior Generation with PMBI

Input: Style s, Simulator interface Sim, Simulation step t, Background AD model $\mathcal{M}_{AD}$, BEV input $\mathcal{X}_{\text{BEV}}$
Output: Style-aligned driving decision D_t

$ctx_t \leftarrow$ Sim.GetContext(t)
if $t == 0$ **then**
  $\mathcal{B}_{desc} \leftarrow$ LLM_Generate(s)
  $\{p_1, p_2, p_3\} \leftarrow$ LLM_Translate(Init, $\mathcal{B}_{desc}$)
end if
if t is the update time of physiological factors **then**
  $p_2 \leftarrow$ LLM_Translate(Update, p_2)
end if
if t is triggered randomly for attentional factors **then**
  $p_3 \leftarrow$ LLM_Translate(ReInterpret, $\mathcal{B}_{desc}$)
end if
$\mathcal{O}_t \leftarrow$ Identify_Objects($\mathcal{X}_{\text{BEV}}$), scripts $\leftarrow$ []
for $o \in \mathcal{O}_t$ **do**
  $APIs \leftarrow$ PMBI_MatchAPIs($o, \{p_1, p_2, p_3\}$)
  compute adjustment parameters for $APIs$ based on last-time values to maintain consistency
  scripts $\leftarrow$ [scripts, $APIs$]
end for
$\mathcal{X}'_{\text{BEV}} \leftarrow$ adjust $\mathcal{X}_{\text{BEV}}$ by scripts
$D_t \leftarrow \mathcal{M}_{AD}(\mathcal{X}'_{\text{BEV}})$
return D_t

	Style-Homogeneous Traffic			Tested AD		Style-Homogeneous Traffic			Tested AD	
	L1	L2	L3	DS	RC	L1	L2	L3	DS	RC
One SIL influence:	normal	normal	normal	100	100	normal	*drunk*	normal	↓52.3%	61.2
	aggressive	normal	normal	↓**54.5%**	100	normal	*fatigued*	normal	↓38.1%	100
	cautious	normal	normal	↓21.9%	100	normal	normal	*distracted*	↓35.4%	100
Two SIL influence:	normal	*drunk*	*distracted*	↓31.3%	100	normal	*fatigued*	*distracted*	↓60.0%	89.0
	aggressive	normal	*distracted*	↓62.7%	86.1	*cautious*	normal	*distracted*	↓48.3%	86.2
	aggressive	*drunk*	normal	↓**64.7%**	100	*cautious*	*drunk*	normal	↓14.7%	100
	aggressive	*fatigued*	normal	↓37.3%	100	*cautious*	*fatigued*	normal	↓37.0%	94.3
Three SIL influence:	*aggressive*	*drunk*	*distracted*	↓**57.0%**	100	*cautious*	*drunk*	*distracted*	↓24.1%	100
	aggressive	*fatigued*	*distracted*	↓55.3%	100	*cautious*	*fatigued*	*distracted*	↓46.7%	94.6

Table 1: Performance of InterFuser AD Model in Style-Homogeneous Traffic. "↓" indicates performance drop from the baseline (first row, separated by a horizontal line). Driving Score (DS) and Route Compliance (RC)

Experiments

Implementation

All simulations are conducted in CARLA 0.9.10 (Dosovitskiy et al. 2017) within Town05, using 30 concurrent background vehicle agents per run. These agents are implemented using the Roach expert policy (Zhang et al. 2021). To assess the impact of style influences, we validate representative traits across three style influence layers: L1 includes *aggressive* and *cautious* personality; L2 models physical traits with episodic (*drunk*) and incremental (*fatigued*) patterns; L3 simulates transient attentional decline (*distracted*). Stable traits update every 2000 simulation steps, while unstable traits are triggered stochastically via a Poisson process with an arrival rate of 0.064. Each experiment contains 10 routes, repeated three times to ensure statistical robustness.

Experimental Setup

Hardware. Rendering is performed on one NVIDIA RTX 4090 GPU. The simulation framework and multi-agent runtime are distributed across eight NVIDIA 2080 Ti GPUs. For language model inference, we use a hybrid configuration: LLaMA 3.1 is deployed locally on an A800 GPU for low-latency validation, while GPT-4o-mini is used for high-level behavioral reasoning and influence script generation.

Metrics. We adopt standard metrics from the CARLA Leaderboard v1 (Dosovitskiy et al. 2017), including *Driving Score* (DS) and *Route Compliance* (RC), to evaluate critical driving safety over complete routes in AD testing.

Baselines. The conventional style-agnostic traffic simulation with a setting (normal, normal, normal) serves as the baseline test environment.

Research Question. We evaluate the effectiveness, realism, and efficiency of `HDSim` by addressing the three questions:

RQ1. Can `HDSim` simulate diverse, style-rich human driving behaviors in traffic and effectively expose the weaknesses of AD models under these conditions?

RQ2. Do the AD failures identified by `HDSim` correspond to real-world incidents, and how realistic are the generated stylized driving behaviors?

RQ3. Does `HDSim` ensure robust stylized driving to support more reliable AD testing, while maintaining system scalability and efficiency for large-scale deployment?

Exposing AD Weaknesses in Stylized Traffic

To answer **RQ1**, we use `HDSim` to construct three types of stylized traffic scenarios: *style-homogeneous*, *style-heterogeneous*, and *selected challenging stylized* traffic.

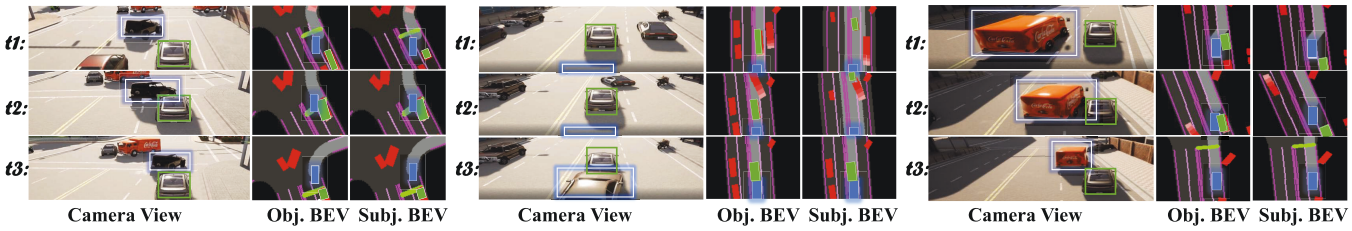
In *style-homogeneous* traffic (see Table 1), all agents in a scenario share the same driving style, defined by different combinations of SILs, including one, two, or all three layers. *style-heterogeneous* traffic includes three styles, each involving one personality trait (normal, *aggressive*, or *cautious*), to represent a distinct type of human driver. These styles are further paired with their corresponding highest-risk outer-layer influences to ensure high-risk diversity and comprehensive style coverage. Each style controls 10 background vehicles in the simulation. The *selected challenging stylized* traffic (see Table 2) set consists of three representative scenarios: the highest-risk single-SIL and highest-risk multi-SIL homogeneous settings, as well as the heterogeneous setting. These are used to evaluate AD model performance under particularly adverse conditions.

We first evaluate the effectiveness of stylized traffic using InterFuser (Shao et al. 2023), which achieves the strongest CARLA performance and serves as the representative AD model. To further explore the generalizability of `HDSim` across model types, we test five additional state-of-the-art AD models: *closed-loop models* include TFPP (Jaeger, Chitta, and Geiger 2023), AIM (Jaeger, Chitta, and Geiger 2023); *open-loop models* include VAD (Jiang et al. 2023) and ST-P3 (Hu et al. 2022); and the *LLM-assisted model* is LMDrive (Shao et al. 2024).

Diverse Traffic Effectiveness in AD Tests. All stylized traffic scenarios result in significantly more hidden failures than the non-stylized baseline (Table 1), demonstrating that diverse human-style behaviors are effective in revealing AD weaknesses. Within *style-homogeneous* traffic (Table 1), the simulated traits and their combinations closely match both the framework design and observed real-world behavioral patterns. The lower-level style traits produce more consistent risk patterns for AD models, with L1 personality traits exerting stronger effects than L2 physiological states and L3 attentional traits. When combined, these styles interact in reinforcing, neutralizing, or interfering ways, showing that style composition is far from a simple linear superposition. For example, drunk tendencies may amplify aggressive behav-

	Selected Challenging Stylized Traffic			TFPP		AIM		VAD		ST-P3		LMDrive	
	L1	L2	L3	DS	RC	DS	RC	DS	RC	DS	RC	DS	RC
Homo.	normal	normal	normal	90.6	100	82.4	100	13.9	34.6	5.56	72.6	19.8	26.4
	aggressive	normal	normal	↓51.6%	94.8	↓67.6%	100	↓26.6%	30.5	↓37.1%	56.0	↓21.9%	22.3
	aggressive	*drunk*	normal	↓54.5%	95.8	↓63.2%	100	↓33.1%	30.1	↓31.7%	55.7	↓23.3%	25.4
Heter.	normal *aggressive* *cautious*	*fatigued* *drunk* normal	*distracted* normal *distracted*	↓29.2%	100	↓24.5%	100	↓28.1%	32.0	↓25.9%	67.0	↓21.7%	25.6

Table 2: Performance of Diverse AD Models in Selected Challenging Stylized Traffic. Homogeneous (Homo.): one style controls 30 vehicles. Heterogeneous (Heter.): Three distinct styles control 10 vehicles each.



(a) Case 1: (*aggressive*, normal, normal) (b) Case 2: (normal, *fatigued*, normal) (c) Case 3: (normal, normal, *distracted*)

Figure 5: Case studies of AD failures involving stylized human drivers. Green boxes: tested AD vehicle, blue-white boxes: driver agents with specific styles, and red boxes: other items. Obj. BEV: Objective BEV; Subj. BEV: Subjective BEV.

ior, causing shorter following, or offset the heavy braking of cautious drivers, whereas distracted traits often disrupt aggressive or drunk styles and lead to unstable behavior. Interestingly, *style-heterogeneous* traffic induces only moderate effects (DS=↓36.9%, RC=87.3), likely due to offsetting interactions among conflicting styles, which lead to more balanced traffic dynamics.

Evaluating AD Models under Challenging Traffic. As shown in Table 2, the *selected stylized challenging* traffic scenarios effectively stress-test all AD models, exposing failure modes often overlooked in conventional settings. Among them, closed-loop models achieve the highest overall performance but show the most limited adaptability under style perturbations. In contrast, open-loop/LLM-assisted models, despite their lower overall performance, maintain greater stability across diverse styles. This may be attributed to their broader consideration of interaction dynamics in model design or the benefits of LLM-guided reasoning.

Realistic Stylized Accidents and Behaviors

To answer **RQ2**, we investigate the realism of style-induced AD accident cases by aligning them with real-world reports. We further validate the similarity between simulated stylized behaviors and real-world driving styles, and compare our results against realism baselines.

Style-Induced Failures: Real Case Analyses. We present three representative style-induced accident cases, each corresponding to a single style influence from L1 to L3. In Case 1, an aggressive driver performs a sudden cut-in without yielding, while the AD model fails to anticipate and continues straight, resulting in a collision (Figure 5a). In Case 2, a fatigued driver reacts late due to cognitive inertia and collides with an abruptly braking AD vehicle (Figure 5b). In Case 3, a distracted driver misses early cues due to degraded perception updates, and the AD model fails to brake in time, leading to a crash (Figure 5c).

L1	L2	L3	CARLA	ProSim	**HDSim(Ours)**
aggressive	normal	normal	82.6	86.5	**97.8**
cautious	normal	normal	66.7	88.3	**98.1**
normal	*distracted*	normal	–	0.00	**45.0**
normal	*fatigued*	normal	–	53.5	**72.7**

Table 3: F1-score (%) of different driving styles evaluated using the mean of RF, SVM, and KNN classifiers.

The failure cases show strong alignment with real-world incidents (Caseid: 2005009501684, 2005041508481, 2007045403168) documented in NHTSA reports (National Highway Traffic Safety Administration 2024), validating the realism of style-induced AD failures. This suggests that the latent failures identified in Tables 1–2 are not artifacts of simulation but likely to occur in real traffic, highlighting HDSim's effectiveness in revealing critical AD blind spots without physical deployment.

Validation of Simulated Style Realism. To validate style realism in HDSim, five drivers (2–3 years' experience) annotated 3,901 style-unlabeled trajectories from the behavior-rich INTERACTION dataset (Zhan and et al. 2019) into four single-SIL driving styles: (*aggressive*, normal, normal), (*cautious*, normal, normal), (normal, *fatigued*, normal), and (normal, normal, *distracted*). Three classifiers (RF, SVM, KNN) were then trained on these annotations to assess the alignment between simulated and real styles, with each style evaluated on 900 simulated trajectories.

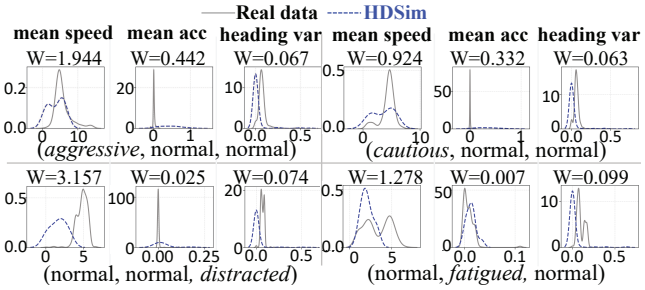


Figure 6: Real-world style alignment of HDSim. Each subplot shows KDE curves comparing real vs simulated trajectories under different driver styles. The X-axis denotes the value of each feature (e.g., speed, acceleration (acc)), and the Y-axis shows the probability density. A smaller Wasserstein distance W implies better alignment.

Tested Stylized Driver			Performance	
L1	L2	L3	DS	RC
normal	normal	normal	91.2	100
aggressive	normal	normal	↓12.0	100
cautious	normal	normal	↓3.20	100
normal	*drunk*	normal	↓12.6	100
normal	*fatigued*	normal	↓8.80	100
normal	normal	*distracted*	↓7.00	100

Table 4: Robust driving evaluation under style diversity.

Realism Baseline. (1) CARLA's parameter-based style module, supporting only (*aggressive*, normal, normal) and (*cautious*, normal, normal) styles; (2) ProSim, which simulates only concrete behavior instructions, not abstract styles. For comparison, we use HDSim-generated stylized behavior descriptions as instruction inputs for ProSim.

As shown in Table 3, HDSim achieves up to 98.1% classification F1-score, outperforming the baselines by an average of 23.3% over CARLA and 21.3% over ProSim. Figure 6 further demonstrates that HDSim exhibits minimal behavioral deviation from real-world driver styles (e.g., mean_acc Wasserstein Distance (Villani et al. 2008), $W = 0.007$ for the (normal, *fatigued*, normal)). Moreover, compared to ProSim, HDSim achieves lower Wasserstein distances in speed (↓38.16%), acceleration (↓27.42%), and heading (↓14.4%), indicating superior realism.

Driver Robustness and System Performance

Robust Decision-Making under Style Diversity. For **RQ3**, we evaluate the performance of driver agents under five single-SIL driving styles within baseline traffic scenarios. As shown in Table 4, each driver preserves robust driving performance while displaying distinct style traits, demonstrating that diverse style influences do not compromise driving robustness. This also rules out insufficient driver robustness as a confounding factor in the AD test results (Tables 1–2). Notably, stylized drivers can achieve any driving capability (e.g., baseline performance) without model modification, enabled by HDSim.

Agent Numbers	GPU (GB/%)	CPU (GB/%)	Driver Sim Step (s)	Rule Sim Step (s)
30	1.211 / 8%	20.4 / 1.3%	0.0229	0.0057
50	1.427 / 10%	22.4 / 1.3%	0.0341	0.0080
70	1.680 / 13%	24.0 / 1.3%	0.0480	0.0131

Table 5: System performance at different agent scales.

Evaluation of Simulation Efficiency and Scalability. For system performance in **RQ3**, runtime overhead remains minimal—CPU/GPU usage scales linearly with 30–70 driver agents, and the local LLM uses only 7.15 GB of GPU memory. LLM reasoning, triggered just 3–6 times per route, is latency-sensitive but fully parallelized. Overall runtime (with LLM inference) remains comparable to standard rule-based CARLA simulations. Full performance metrics are reported in Table 5 (1 2080 Ti, 12 GB GPU).

Related Work

Driving Behavior Simulation in Traffic. Existing methods fall into: *log-playback*, *rule-based*, and *learning-based*. Log-playback replays recorded trajectories but lack interactivity (Montali et al. 2023; Li et al. 2022). Rule-based approaches offer interpretability but limited behavioral diversity (Behrisch et al. 2011; Dosovitskiy et al. 2017). Learning-based methods improve realism through imitation or reinforcement learning (Schulman et al. 2017), yet rely on large labeled datasets and struggle to generalize to rare scenarios (Suo et al. 2021; Zhang et al. 2025). Recent LLM-based methods, such as ProSim (Tan et al. 2024), offer customizable behavior through fine-tuning and explicit instructions. However, they remain limited to low-level vehicle action simulation and do not model high-level driver traits.

Driver Style Simulation. Existing methods fall into two categories: *parameter-based*, which manually tune behavioral parameters (e.g., aggressive speed) but lack scalability (Dosovitskiy et al. 2017; Niehaus and Stengel 1991); and *LLM-finetuned*, which generate style-specific behaviors using fine-tuned LLMs (Yang et al. 2024) but incur high costs and unstable real-time performance. Both are typically limited to a few predefined styles (only aggressive, cautious) and often implement them in a rigid, fragmented manner.

Conclusion

We present HDSim, a simulation framework that models diverse human-like driving styles in background traffic for AD evaluation. Unlike rule-based or low-level models, it represents driving styles as layered cognitive influences that bias perception, enabling style-specific decisions through the PMBI mechanism. PMBI uses LLM-generated programs to inject perception bias without altering AD models, supporting scalable and stable simulation. Experiments show that HDSim uncovers more hidden issues than conventional testing by providing realistic human-driven traffic risks without physical deployment. Nonetheless, limitations remain, such as the formalization of the style model, and we will continue advancing HDSim through future research.

Acknowledgments

This work was supported in part by NSFC under Grant 62272224, Grant 62572228, Grant 61872176, Grant 62272215, Grant 62302207, and Grant 62432004; in part by the Leading Edge Technology Program of the Jiangsu Natural Science Foundation under Grant BK20202001; in part by the Science Foundation for Youths of Jiangsu Province under Grant BK20220772; and in part by Nanjing University-China Mobile Communications Group Company Ltd., Joint Institute.

References

Behrisch, M.; Bieker, L.; Erdmann, J.; and Krajzewicz, D. 2011. SUMO–simulation of urban mobility: an overview. In *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*. ThinkMind.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Elander, J.; West, R.; and French, D. 1993. Behavioral correlates of individual differences in road-traffic crash risk: An examination of methods and findings. *Psychological bulletin*, 113(2): 279.

Endsley, M. R. 2017. Toward a theory of situation awareness in dynamic systems. In *Situational awareness*, 9–42. Routledge.

Green, T. R. G.; and Petre, M. 1996. Usability analysis of visual programming environments: a 'cognitive dimensions' framework. *Journal of Visual Languages & Computing*, 7(2): 131–174.

Hu, S.; Chen, L.; Wu, P.; Li, H.; Yan, J.; and Tao, D. 2022. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, 533–549. Springer.

Jaeger, B.; Chitta, K.; and Geiger, A. 2023. Hidden biases of end-to-end driving models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8240–8249.

Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8350.

Kahneman, D. 1973. *Attention and effort*, volume 1063. Citeseer.

Li, Q.; Peng, Z.; Feng, L.; Zhang, Q.; Xue, Z.; and Zhou, B. 2022. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(3): 3461–3475.

Montali, N.; Lambert, J.; Mougin, P.; Kuefler, A.; Rhinehart, N.; Li, M.; Gulino, C.; Emrich, T.; Yang, Z.; Whiteson, S.; et al. 2023. The waymo open sim agents challenge. *Advances in Neural Information Processing Systems*, 36: 59151–59171.

National Highway Traffic Safety Administration. 2024. NMVCCS XML Case Viewer. https://crashviewer.nhtsa.dot.gov/LegacyNMVCCS/Search. Accessed on August 2, 2025. Provides case-level query interface for the National Motor Vehicle Crash Causation Survey (NMVCCS).

Niehaus, A.; and Stengel, R. F. 1991. An expert system for automated highway driving. *IEEE Control Systems Magazine*, 11(3): 53–61.

Norman, D. A.; and Shallice, T. 1986. Attention to action: Willed and automatic control of behavior. In *Consciousness and self-regulation: Advances in research and theory volume 4*, 1–18. Springer.

Philip, P.; Vervialle, F.; Le Breton, P.; Taillard, J.; and Horne, J. A. 2001. Fatigue, alcohol, and serious road crashes in France: factorial study of national data. *Bmj*, 322(7290): 829–830.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Schwarz, K. A.; and Weller, L. 2023. Distracted to a fault: Attention, actions, and time perception. *Attention, Perception, & Psychophysics*, 85(2): 301–314.

Shao, H.; Hu, Y.; Wang, L.; Song, G.; Waslander, S. L.; Liu, Y.; and Li, H. 2024. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15120–15130.

Shao, H.; Wang, L.; Chen, R.; Li, H.; and Liu, Y. 2023. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, 726–737. PMLR.

Sheeran, P.; Orbell, S.; and Trafimow, D. 1999. Does the temporal stability of behavioral intentions moderate intention-behavior and past behavior-future behavior relations? *Personality and Social Psychology Bulletin*, 25(6): 724–734.

Suo, S.; Regalado, S.; Casas, S.; and Urtasun, R. 2021. Trafficsim: Learning to simulate realistic multi-agent behaviors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10400–10409.

Tan, S.; Ivanovic, B.; Chen, Y.; Li, B.; Weng, X.; Cao, Y.; Krähenbühl, P.; and Pavone, M. 2024. Promptable closed-loop traffic simulation. *arXiv preprint arXiv:2409.05863*.

Taubman-Ben-Ari, O.; Mikulincer, M.; and Gillath, O. 2004a. The multidimensional driving style inventory—scale construct and validation. *Accident Analysis & Prevention*, 36(3): 323–332.

Taubman-Ben-Ari, O.; Mikulincer, M.; and Gillath, O. 2004b. The multidimensional driving style inventory—scale construct and validation. *Accident Analysis & Prevention*, 36(3): 323–332.

Villani, C.; et al. 2008. *Optimal transport: old and new*, volume 338. Springer.

Yang, R.; Zhang, X.; Fernandez-Laaksonen, A.; Ding, X.; and Gong, J. 2024. Driving style alignment for llm-powered driver agent. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11318–11324. IEEE.

Zhan, W.; and et al. 2019. INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps. *arXiv preprint arXiv:1910.03088*.

Zhang, Z.; Jia, X.; Chen, G.; Li, Q.; and Yan, J. 2025. TrajTok: Technical Report for 2025 Waymo Open Sim Agents Challenge. Technical report, Shanghai Jiao Tong University. Equal contributions by Zhiyuan Zhang and Xiaosong Jia. Corresponding author: Junchi Yan.

Zhang, Z.; Liniger, A.; Dai, D.; Yu, F.; and Van Gool, L. 2021. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15222–15232.