Transportation Research Part C 186 (2026) 105606



Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc





LLM agents can partially mimic human driving behaviors and decision values but better align with aggressive profiles

Pu Wang [a], Guanlin Wu [a], Yang Zhao [a], Yichen Lin [a], Hao Frank Yang [a,b,*]

[a] Department of Civil and Systems Engineering, Johns Hopkins University, Baltimore, MD 21218, United States
[b] Data Science and AI Institute, Johns Hopkins University, Baltimore, MD 21218, United States

ARTICLE INFO

Keywords:
Large language models
Multimodal LLM
Data-centric AI
Human likeness
Driving behavior

ABSTRACT

As autonomous vehicles (AVs) increasingly share roads with human drivers, creating human-like driving agents has emerged as a promising approach for seamless integration into mixed-autonomy environments. Achieving effective vehicle-human interaction is becoming a clear and long-term challenge. A notable direction is leveraging Large Language Models (LLMs) as driving agents, given their human-like reasoning abilities and adaptability to diverse and dynamic driving scenarios. However, a foundational question arise – do LLMs really like human drivers, and to how much content? In this paper, we present a novel framework to evaluate whether LLM- and multimodal LLM-based agents can authentically simulate human driving behavior. To this end, we introduce HermesSense, a benchmark dataset grounded in real-world human driving behavior that integrates both sensory and semantic information. Built upon this dataset, our evaluation methodology assesses behavioral alignment from both short-term decision-making and long-term statistical perspectives, quantifying how closely these models reproduce human-like driving decisions under identical conditions and when conditioned on specific driver profiles, respectively. To further illuminate the factors shaping human-like behavior, we conduct a data attribution analysis to identify which components in the input most influence the LLM agents' behavioral decisions. Thirteen state-of-the-art LLMs from five families were evaluated, revealing notable differences in their ability to mimic human driving behavior. The proposed dataset, benchmark, and feature attribution insights lay a solid foundation for the future optimal integration of information flow in AV and robotics research, fostering a more transparent and reliable human-machine collaborative environment.

1. Introduction

The era of mixed autonomy has arrived. Autonomous Vehicles (AVs) are now being tested and deployed in ride-hailing operations across five cities, making their interaction with human drivers increasingly important (Gao et al., 2024; Wu et al., 2025b). From now on and into the coming years, transportation environments will be characterized by mixed autonomy, requiring AVs to coexist with human drivers. Therefore, an increasing amount of research has focused on better integration with human traffic, including building human-like driving agent (Bhattacharyya et al., 2023; Lee et al.), proposing human-aligned decision-making strategies (Hang et al., 2020; Hu et al., 2024; Wu et al., 2024b), and human-like simulation environments for autonomous vehicles (Zhou et al., 2024; Stead et al., 2022), aiming to make autonomous driving agents behave more like humans for interaction, safety, efficiency improvements.

* Corresponding author.
E-mail address: haofrankyang@jhu.edu (H.F. Yang).

https://doi.org/10.1016/j.trc.2026.105606
Received 22 September 2025; Received in revised form 10 January 2026; Accepted 24 February 2026
Available online 14 March 2026
0968-090X/© 2026 Published by Elsevier Ltd.

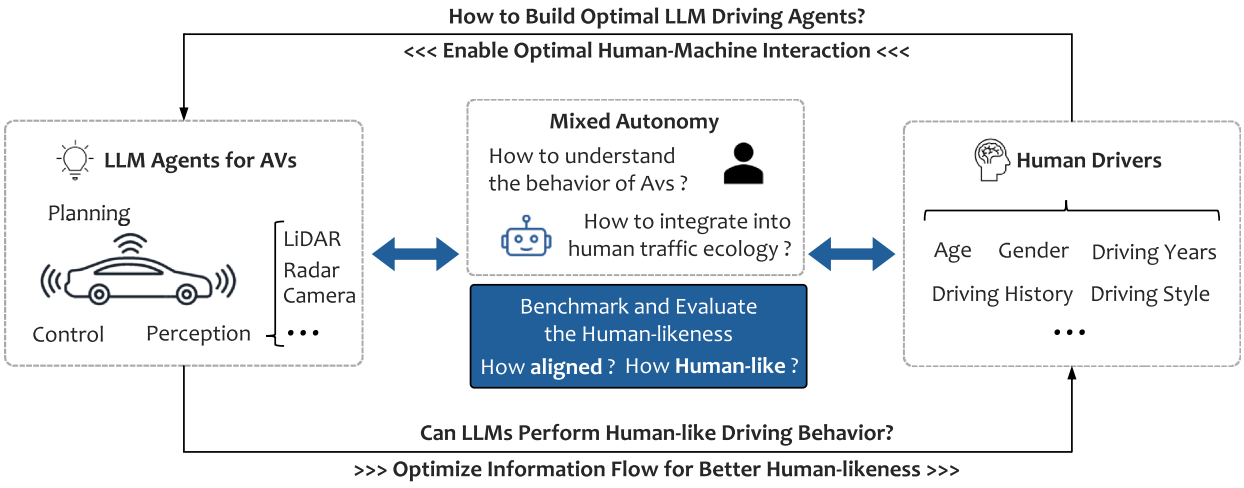


Fig. 1. Current challenges in mixed autonomy and the path toward better human-machine interaction. Mixed autonomy poses challenges in aligning AV behavior with human drivers. Our evaluation and data attribution framework guides the development of interpretable, human-like LLM agents for smoother traffic integration.

However, whether we aim to build human-like driving agents or better human-vehicle interaction, the task is extraordinarily challenging, primarily due to the intrinsic complexity of human nature, as illustrated in Fig. 1. As Dr. Richard Thaler, the 2017 Nobel Prize winner in Economic Sciences, noted in his work on behavioral economics, *"Conventional economics assumes that people are highly rational – super-rational – and unemotional. They can calculate like a computer and have no self-control problems."* (Orrell, 2010; Barberis and Thaler, 2003) A similar challenge arises in the field of autonomous driving: traditional learning-based driving agents struggle to replicate genuinely human behavior because of several inherent difficulties associated with human drivers. Human driving behavior is highly diverse, shaped by complex cognitive mechanisms influenced by nonlinearity, uncertainty, and randomness (Wang et al., 2014; Zhu et al., 2018). Each driver responds uniquely to similar scenarios, making it challenging to learn and integrate these behaviors into driving agents. This complexity is amplified by the multimodality of driving data, such as sensor-based trajectories (Yang et al., 2025), textual incident reports (Wu et al., 2024a), and video/image recordings (Zhang et al., 2025b), which traditional models struggle to unify into coherent representations. A promising direction is leveraging Large Language Models (LLMs) to develop driving agents that better interact with human drivers (Jin et al., 2024). Utilizing their advanced capabilities in human-like text gexneration and reasoning, LLMs provide a novel but risky paradigm for decision-making within AVs (Sha et al., 2025). LLMs are extensive deep learning models comprising billions of parameters, trained on vast textual datasets. Beyond generating highly convincing human-like text, these models exhibit remarkable emerging abilities, including solving analogical reasoning tasks (Musker et al., 2024; Zhao et al., 2025a,b), programming complex applications (Nam et al., 2024), employing tools for multitask scenarios (Shen, 2024), and dynamically adapting strategies based on contextual cues (Yang et al., 2023a). Thus, several critical questions arise: **Do LLM agents perform like human drivers? If so, to what extent? Which agent performs most similarly to humans?**

If the answer is yes, the next step is to understand why and how – which parts of the input data empower the LLMs' acting like human drivers and how different LLMs respond to and interpret these human priors? Given LLMs', especially multimodal LLMs' (MLLMs) ability to integrate diverse forms of driving-related information in prompts (Islam, 2024), it becomes essential to understand how different inputs shape the agent's behavior (Chen and Lu, 2024). Feature attribution frameworks are employed to quantify the contribution of each input component to the model's output. By analyzing feature attributions, we gain deeper insights into how LLM agents reason and which types of information they rely on when mimicking human drivers. Such understanding can inform more targeted data collection strategies for developing human-like LLM-based agents and guide the design of pretraining objectives that better align with human decision-making (Lee et al., 2025; Chang et al., 2024).

To answer these critical questions, we first build a comprehensive human driving dataset and benchmark that encompasses truly comprehensive information about real drivers: demographic backgrounds, historical crash records, real-time observed driving behaviors, real-time vehicle trajectories, speed, string wheel angles, and even real-time driver gaze attention masks, and etc (Geyer et al., 2020). This diverse dataset forms the cornerstone of our study, providing a nuanced view of both typical and individual-specific driving patterns. Next, we use expert-guided textualization and reformatting techniques to convert the recorded human driving decisions into multimodal prompts. These prompts capture key elements of the drivers' decision-making processes, ranging from environment cues to personal profiles, and serve as inputs to LLMs and MLLMs (Han et al., 2025; Li, 2024). Our goal is to determine how accurately these models can interpret and reproduce human-like decisions in various driving scenarios.

To rigorously evaluate the "human-likeness" of decisions made by LLM-based agents in driving tasks, we measure their behavioral alignment from both short-term and statistical (long-term) perspectives (Vinayavekhin et al., 2017). Firstly, we quantify whether the agents' decisions match those of human drivers under identical driving conditions, thereby assessing the models' short-term alignment with human behavior. Furthermore, we investigate whether these agents, when conditioned on specific driver profiles

Table 1
Comparison of well-known existing AV datasets. *R* and *S* indicate whether the dataset is real-world or synthetic. Frames/Video/Text under **Size** refer to the number of image frames, video duration (h: hours, v: video clips, c: clips), and word counts. `pl` for planning, `perc` for perception, and `pl & behav` for planning tasks for driver behavior. The ✓ indicates the presence of the corresponding modality, and the ○ marks manually labeled driver behavior.

Dataset	Task	Size	Image/Video		Sequence Data		Text Data				Venue
		Frames/Video/Text	Driving Scenes	Driver Gaze	Vehicle Trajectory	Vehicle Dynamics	Envir-onment	Traffic Dynamics	Driver Behhavior	Driver Info	
BDD100K $_{(R)}$ (Yu et al., 2020)	Perc.	100K / 1.1K (h) / –	✓	–	✓	–	–	–	–	–	CVPR-18
nuScenes $_{(R)}$ (Caesar et al., 2020)	Perc.	40K / 5.5 (h) / –	✓	–	✓	✓	–	–	–	–	CVPR-20
ONCE $_{(R)}$ (Mao et al., 2021)	Perc.	7M / 144 (h) / –	✓	–	✓	–	–	–	–	–	NeurIPS-21
K-Radar $_{(R)}$ (Paek et al., 2022)	Perc.	35K / – / –	✓	–	✓	✓	–	–	–	–	NeurIPS-22
Bench2Drive $_{(R,S)}$ (Jia et al., 2024)	Pl.	2M / 13K (c) / –	✓	-	✓	–	–	–	○	–	NeurIPS-24
LaMPilot $_{(S)}$ (Ma et al., 2024)	Pl.	- / - / 35K	–	–	✓	✓	–	–	○	–	CVPR-24
NuPrompt $_{(R)}$ (Wu et al., 2025a)	Perc.	34K / 850 (v) / 2.5M	✓	–	✓	✓	–	✓	–	–	AAAI-25
DriveLM $_{(R,S)}$ (Sima et al., 2024)	Perc., Pl.	10K / – / 440K (QA)	✓	–	✓	✓	✓	–	○	-	ECCV-24
VLAAD $_{(R)}$ (Park et al., 2024)	Pl.	- / 11K (v) / 3M	✓	-	✓	✓	✓	✓	✓	–	WACV-24
HermesSense$_{(R)}$	Pl. & Be-hav.	2.5M / 24 (h) / 10M	✓	✓	✓	✓	✓	✓	✓	✓	–

based on accumulated decision records (e.g., a cautious driver with a history of minor collisions), replicate the characteristic decision patterns thus gauging their ability to capture the statistical variations over the long term. Following these evaluations, we conduct a systematic analysis of the prompt data attributes. Our findings highlight the critical components, such as demographic indicators, gaze attention, and context-specific driving cues, that most heavily influence the models' ability to generate human-like decisions under various traffic conditions, and comparing with information importance considered by human drivers (Koorathota et al., 2024). Based on these insights, we offer perspectives on how to refine prompt design and model architectures for even more accurate behavioral modeling. By establishing human-likeness evaluation criteria and identifying pivotal information sources within the input data, our framework paves the way for future advancements in human behavior understanding, enhanced human-machine interaction, and developing human-like LLM agents in autonomous driving systems.

2. Related work

Vision and LiDAR datasets for AV perception and planning. Existing datasets such as Waymo (Sun et al., 2020), nuScenes (Caesar et al., 2020), ONCE (Mao et al., 2021), and Bench2Drive (Jia et al., 2024) have advanced AV perception and planning tasks by providing extensive visual and multimodal sensor data. Utilizing these datasets, recent studies have demonstrated deployable performance in crucial tasks, including object detection, tracking, trajectory prediction, and behavioral and motion planning (You et al., 2024; Liu et al., 2026). Currently, most testing scenarios occur within mixed-autonomy environments, where AVs must coexist and interact with human drivers. Integrating AVs into the human-driven traffic ecology while minimizing discrepancies remains a significant challenge (Chen et al., 2024b). Recent efforts have attempted to bridge this gap by incorporating textual semantics, providing richer contextual understanding, and improving interpretability in AV decision-making (Ma et al., 2024; Wu et al., 2025a; Sima et al., 2024; Park et al., 2024; Zhang et al., 2025a,c; Hou et al., 2025; Liu et al., 2025b). However, as shown in Table 1, existing datasets lack explicit features capturing mixed-autonomy interactions, especially detailed comparisons between AV behaviors and human driving behaviors, as well as their interactions. These features are critical for modeling human-like decision-making in mixed-autonomy scenarios (Hu et al., 2022).

LLM-Agent for Human-like AV planning and decision making. LLMs have exhibited remarkable progress in linguistic understanding and reasoning, prompting extensive research into their potential for enhancing autonomous driving (Yang et al., 2023b). Recent efforts have examined natural language-based perception, LLM-enabled human-like decision making, and the development of LLM-driven agents (Long et al., 2024; Wang et al., 2023). For instance, Long et al. (2024) and Wang et al. (2023) leveraged multimodal LLMs to refine vehicle control and bolster decision-making in mixed-autonomy scenarios. Concurrently, Cui et al. (2023) incorporated LLMs into conventional driving stacks, capitalizing on commonsense reasoning to improve the robustness of AV decisions. A growing line of work focuses specifically on emulating and interacting with human behavior in real-world driving contexts. For example, Yang et al. (2024) proposed a multi-alignment framework that aligns LLM-driven agents with human driving styles by using data gathered from post-driving interviews. Their findings underscore the importance of high-quality human data, such as demographic and behavioral

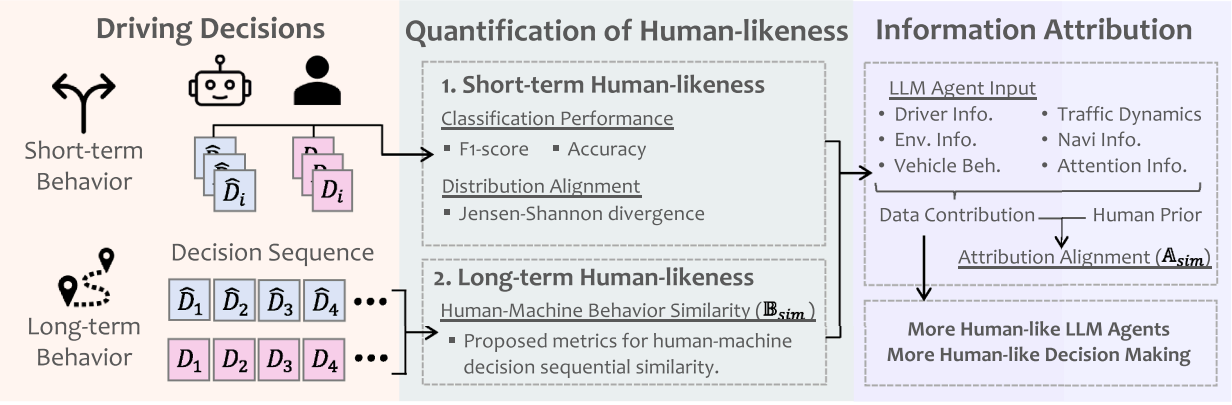


Fig. 2. Introduction to the human-likeness evaluation framework. This framework evaluates human-likeness from both short-term and long-term perspectives, capturing performance variations across diverse driving scenarios and over extended periods. Through the application of data attribution, it also reveals the information reliance patterns of humans and machines, providing insights to optimize information flow and enhance human-machine collaboration.

information in achieving truly human-like autonomous driving behaviors (Mao et al., 2023; Jin et al., 2023). Nonetheless, besides the remaining questions raised in the introduction, much of the existing research has been confined to simulations or datasets that lack comprehensive real-world driver data, limiting both the accurate evaluation of human-likeness and the capacity to learn genuine human behaviors (Kujanpää et al., 2024; Rock et al., 2023).

3. Driving decision agent and human alignment

The human-likeness of LLM-based agent can be assessed through two dimensions, as shown in Fig. 2: **1) Short-term alignment with human drivers** evaluates whether the agent's choices align with human behavior in specific driving contexts (Yang et al., 2024), which is a fundamental requirement for safe integration into mixed-autonomy environments. **2) Long-term (Statistical) likeness to human drivers** measures the temporal consistency of the agent's decision patterns and their alignment with distinct human driving profiles (Ou et al., 2024). More specifically, short-term alignment focus on comparing the agent's discrete driving decisions such as turn left to human's decision given the same input context. On the other hand, long-term human-likeness mainly evaluates whether the agent's temporally extended decision patterns are consistent with and align with distinct human driving profiles over time (e.g., characteristic patterns across a trip). This human-likeness is essential for enabling personalized driving and offers the potential to generate synthetic data representing underrepresented driver populations (Sun et al., 2024). As illustrated in Fig. 3, given a textual prompt x_i from the dataset, we evaluate these two dimensions of human-likeness across three decision-making tasks: `speed` s_i, `turning` t_i, and `steering` e_i. We denote the LLM agent's response as an action tuple $a_i = (s_i, t_i, e_i)$ conditioned on the input prompt x_i. In the following sections, we detail the evaluation metrics and describe the construction of the human-centric dataset used to support this analysis.

3.1. Behavioral alignment with human divers

LLM agent decision alignment metrics. We treat short-term prediction as a context-aware classification problem across three decision-making tasks. The model selects one discrete action per task and generates an explanation grounded in the input scenario. To evaluate the model's behavioral alignment with human drivers, we introduce two metric classes targeting different aspects of this alignment:

Classification performance. We compute standard classification metrics accuracy (Acc.) and F1-score (F1) (Sokolova and Lapalme, 2009) to assess correctness between model predictions and human behaviors. Given the standard notations: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), these metrics are defined as:

- **Accuracy (Acc.)**: Measures the proportion of correctly classified instances over all predictions (Sokolova et al., 2006):

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **F1-score (F1)**: Harmonic mean of correct positive predictions and missed detections, emphasizing performance on the positive class (Goutte and Gaussier, 2005):

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{2}$$

Distributional alignment. Beyond individual decisions, we assess task-level behavior by comparing predicted and human action distributions using the Jensen-Shannon divergence (JSD). Given a set of human behavior instances, we denote the aggregated predicted action distribution as $\hat{p} = (\hat{p}^s, \hat{p}^t, \hat{p}^e)$, and the corresponding ground-truth distribution as $p = (p^s, p^t, p^e)$, where each component represents the empirical distribution over speed (s), turning (t), and steering (e) actions, respectively. We define distributional alignment score $\mathbb{D}_{\text{ali}}$ as:

$$\mathcal{J}(p \parallel \hat{p}) = \frac{1}{2}\mathcal{KL}(p \parallel m) + \frac{1}{2}\mathcal{KL}(\hat{p} \parallel m), \quad m = \frac{1}{2}(p + \hat{p}), \tag{3}$$

where $p \in \{p_T^s, p_T^t, p_T^e\}$ denotes the empirical distribution for each sub-task, and $\mathcal{KL}(\cdot \parallel \cdot)$ denotes the Kullback-Leibler divergence.

Metrics of information attribution for human-likeness. We employ Shapley values to quantify the contribution of various information inputs to the LLM agents' human-likeness (Shapley, 1953). Unlike traditional machine learning models that primarily handle fixed-length feature vectors, LLMs process variable-length text sequences as input (Chen et al., 2023). This characteristic makes commonly used Shapley value approximation methods, such as KernelSHAP (Lundberg and Lee, 2017) and DeepSHAP, less applicable to LLMs. Recent approaches like TokenSHAP (Goldshmidt and Horovicz, 2024) and TransSHAP (Kokalj et al., 2021) have been proposed to address this by decomposing input text into tokens and computing Shapley values at the token level. However, applying token-level Shapley value computation to our tested LLMs introduces two primary challenges: 1) Computational limitations. The computational complexity of Shapley values is exponential in the number of players. In our setting, with an input size of approximately 300 tokens, large-scale computation of token-level Shapley values for driving data becomes impractical. 2) Limited interpretability. Decomposing the prompt at the token level disregards inter-token dependencies, and the arbitrary masking or replacement of tokens can lead to semantic ambiguity and contextual shifts.

In light of these challenges, we propose a paragraph-level Shapley value attribution method which maintains semantic integrity and preserves inter-token dependencies by treating each paragraph as a coherent information unit. Specifically, as detailed in Section 5.1 of the modified manuscript, each driving scene prompt x_i in the dataset is partitioned into $n = 6$ disjoint parts, corresponding to n players. These players represent six information categories: c_0: *driver demographic information*, c_1: *environmental information*, c_2: *vehicle dynamics*, c_3: *traffic dynamics*, c_4: *navigation information*, and c_5: *driver attention information*. To quantify the contribution of each category, we treat them as players in the Shapley value computation. For any given subset of indices $S \subseteq \{0, 1, \ldots, n-1\}$, we construct a filtered input prompt $x_i(S)$ by applying a masking operation on the original segments. Let $x_{i,j}$ denote the text segment of prompt x_i corresponding to category c_j. The constructed prompt is defined as:

$$x_i(S) = \underset{j=0}{\overset{n-1}{\text{concat}}} \left(\mathbb{I}[j \in S] \cdot x_{i,j} \right) \tag{4}$$

where concat$(\cdot)$ is the text concatenation operation, and $\mathbb{I}[\cdot]$ serves as a selection function. Specifically, the term $\mathbb{I}[j \in S] \cdot x_{i,j}$ retains the content of segment $x_{i,j}$ if the category index j is present in S; otherwise, it returns an empty string $\varnothing$. This mechanism ensures that the model input $x_i(S)$ solely reflects the information represented by the coalition S.

The resulting test dataset based on S is defined as $\mathcal{D}(S) = \{x_i(S) | i = 0, 1, ..., M\}$, where M is the size of the testset. We then define the task-specific utility function $v_k(\mathcal{D})$ with $k \in \{s, t, e\}$ denoting to the three driving tasks: *speed* (s), *turning* (t), and *steering* (e). Here, $v_k(\mathcal{D})$ represents the performance metric (e.g., accuracy) on dataset $\mathcal{D}$ for task k. The Shapley value $\phi_{p,k}$, representing the contribution of information category p to task k, is computed as:

$$\phi_{p,k} = \sum_{S \subseteq \mathcal{N} \setminus \{p\}} \frac{|S|! \cdot (n - |S| - 1)!}{n!} \left[v_k(\mathcal{D}(S \cup \{p\})) - v_k(\mathcal{D}(S)) \right], \tag{5}$$

where $\mathcal{N} = \{0, 1, \ldots, 5\}$ is the set of all players.

For each task k, we rank the information parts based on their corresponding contribution to obtain the model-based ranking $\hat{\mathrm{r}}^{(k)}$. The human-perceived importance ranking, derived from expert annotation, is denoted by $\mathrm{r}^{(k)}$. We propose an *attribution similarity* metric $\mathbb{A}_{\text{sim}}$ to evaluate the similarity between the two rankings, defined as the average Kendall's τ correlation (Kendall, 1938) across all tasks:

$$\mathbb{A}_{sim} = \sum_k \tau^{(k)}/n, \quad \tau^{(k)} = (n_c - n_d)/\binom{n}{2}, \tag{6}$$

where $n = 3$ is the number of tasks, and n_c and n_d represent the numbers of concordant and discordant pairs, respectively, among all pairs of information categories (j, j') with $j < j'$. A pair is considered concordant if the relative ordering is consistent in both rankings, i.e., $(\hat{\mathrm{r}}_j^{(k)} - \hat{\mathrm{r}}_{j'}^{(k)})(\mathrm{r}_j^{(k)} - \mathrm{r}_{j'}^{(k)}) > 0$, and discordant if the product is negative.

3.2. Statistical human-likeness

Statistical human-likeness evaluates the alignment between predicted and human driving behaviors over temporally extended sequences. Instead of focusing on isolated decisions, it captures consistency in action patterns and their correspondence to human

Table 2
Semantic cost matrices δ_s, δ_t, and δ_e for speed, turning, and steering control tasks. This table can be put in supplementary.

Speed δ_s				Turning δ_t				Steering δ_e			
GT/Pred	dec.	maintain	acc.	GT/Pred	left	none	right	GT/Pred	left	maintain	right
dec.	0	0.5	1	left	0	0.5	1	left	0	0.5	1
maintain	0.5	0	0.5	none	0.5	0	0.5	maintain	0.5	0	0.5
acc.	1	0.5	0	right	1	0.5	0	right	1	0.5	0

Table 3
Overview of data types and modalities in HermesSense dataset. The HermesSense dataset integrates diverse data modalities including video, numerical signals, and textual descriptions from multiple acquisition sources to enable multimodal analysis of LLM driving agents' behavior.

Data Category	Data Modality	Data Component	Acquisition Source
Driving Scenes Data	Image/Video	Front-view Scene Video	3 × Scene Cameras
Vehicle Dynamic Data	Numerical	Acceleration, Vehicle Speed, Trajectory, Steering Angle, GasPressed, BrakePressed, etc.	CAN bus Loggers, GPS
Driver Demographic	Text	Race, Gender, Commute Frequency, Prescription Glasses, Accident History, etc.	Interview
Driver Gaze Data	Numerical	HeadPos, GazeOrigin, GazeDirection, RightEyePosition, LeftEyePosition, etc.	SmartEye Pro v5.0
Text Description	Text	Environment Description, Navigation Suggestion, Vehicle Dynamics, etc.	Expert-assisted Annotation

driving styles over time. We propose **Human-machine Expected Behavioral Similarity** to quantify this alignment. For each control task and each of the N samples, we define the ground-truth and predicted action sequences as $S = (a_1, a_2, \ldots, a_L)$, $\hat{S} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_L)$. The behavioral similarity score $\mathbb{B}_{\text{sim}}$ is computed using the normalized Levenshtein distance (Levenshtein et al., 1966; Biten et al., 2019):

$$\mathbb{B}_{\text{sim}} = \frac{1}{N} \sum_{j=1}^{N} \left(1 - \mathcal{L}_n(S_j, \hat{S}_j)\right), \tag{7}$$

where $\mathcal{L}_n(S, \hat{S})$ denotes the normalized Levenshtein distance between the ground-truth and predicted sequences, defined as:

$$\mathcal{L}_n(S, \hat{S}) = \frac{1}{\max(|S|, |\hat{S}|)} \sum_{i=1}^{L} C(a_i, \hat{a}_i), \tag{8}$$

with $C(a_i, \hat{a}_i)$ representing the semantic cost of transforming action a_i to $\hat{a}_i$. This cost is composed of three components:

$$C(a_i, \hat{a}_i) = \delta_s(s_i, \hat{s}_i) + \delta_t(t_i, \hat{t}_i) + \delta_e(e_i, \hat{e}_i), \tag{9}$$

where δ_s, δ_t, and δ_e capture the differences in speed, turning, and steering, respectively. For example, in the turning task, replacing a `left` turn with a `right` turn yields a higher cost than substituting it with a `slight right`, reflecting greater behavioral divergence. The detailed cost terms are summarized in Table 2.

4. HermesSense dataset

To evaluate the human-likeness of decision-making in LLM agents, we introduce HermesSense, a human-centric multimodal sensory and semantic dataset grounded in real-world driving behavior. As illustrated in Fig. 3, HermesSense captures the full spectrum of driver perception and action through a richly structured data workflow that includes visual, textual, numerical, and human-centric information. This enables the generation of comprehensive prompts for large language model LLM driving agents. The dataset construction follows a two-stage pipeline: (1) the *collection* of multimodal sensor and human behavioral data including vehicle dynamics, driver gaze, and demographic profiles and (2) the *annotation* of semantic descriptions such as environmental context, navigation instructions, and attention regions. These descriptions are critical for enabling LLMs to reason about driving scenarios from a human-like perspective. A detailed overview of data types is provided in Table 3.

4.1. Dataset collection and preprocess

Dataset collection. The collection of our proposed dataset integrates multiple information streams to comprehensively capture human driving behavior. We recruited 25 drivers of various ages and genders in Southeastern Pennsylvania. Each of them underwent an interview to collect demographic information, shown as Table 4 and was provided with an overview of the driving routes without being informed of specific research objectives to ensure natural driving conditions.

To fully record and analyze their driving behavior, we chose two routes in Manassas, Virginia, shown as below:

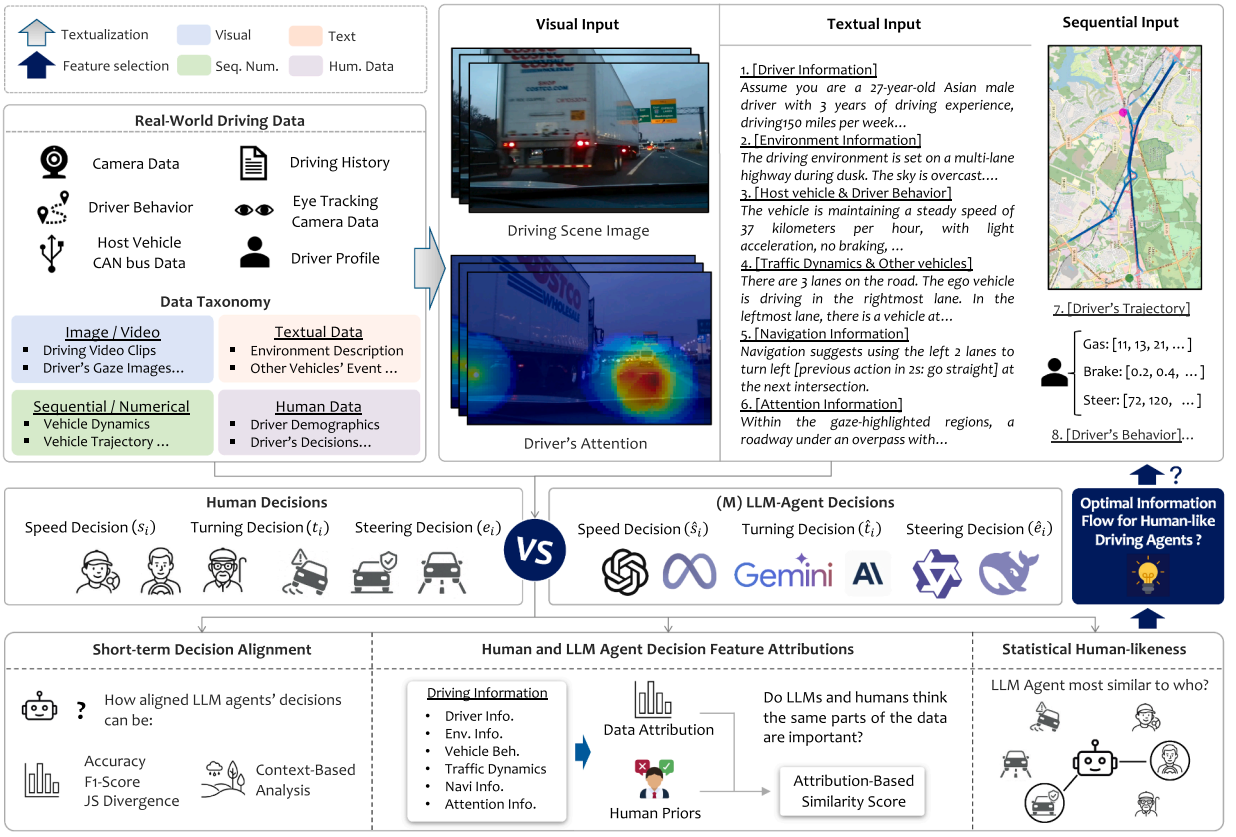


Fig. 3. Overview of proposed HermesSense dataset, benchmark and workflow. HermesSense integrates real-world multimodal driving data, including visual, textual, numerical, and human inputs to generate prompts for LLM-based driving agents. The benchmark evaluates human-likeness through short-term decision alignment, statistical similarity across driver types, and a feature attribution framework for selecting optimal information flow in driving agents.

Table 4
Demographic background related to the participated drivers.

Demographic Category	Category / Group	Count (%)
Total Valid Subjects	Drivers	25
Gender	Male	17 (68%)
	Female	8 (32%)
Race / Ethnicity	Asian	15 (60%)
	White / Caucasian	5 (20%)
	Hispanic / Latino	3 (12%)
	Black / African American	2 (8%)
Age Range	20–25 years	6 (24%)
	26–30 years	9 (36%)
	31–35 years	6 (24%)
	36–45 years	4 (16%)
Driving Experience	1–5 years	12 (48%)
	6–10 years	7 (28%)
	11–20 years	4 (16%)
	21 + years	2 (8%)
Weekly Mileage	<50 miles	6 (24%)
	50–100 miles	8 (32%)
	100–200 miles	9 (36%)
	200 + miles	2 (8%)
Vision Correction	No correction needed	16 (64%)
	Contacts / glasses required	9 (36%)
Accident History	No accidents	19 (76%)
	Previous accidents	6 (24%)
Traffic Violations	No violations	22 (88%)
	Previous violations	3 (12%)

Table 5
Dataset statistics before and after gaze filtering.

	Daytime	Evening
Number of trips	28	20
Number of scene frames	1,535,143	1,001,924
Frames without valid gaze	483,570	320,616
Dropout rate (%)	31.5	32.0
Valid gaze frames	1,051,573	681,308

Table 6
Distribution of driving behaviors across different demographic groups.

	Driving Style			Driver Age		Driving Experience		Driver Gender	
Behavior	Aggressive	Moderate	Defensive	≤25	>25	≤5 yrs	>5 yrs	Male	Female
Accelerate	0.200	0.258	0.247	0.200	0.254	0.241	0.247	0.230	0.260
Maintain speed	0.545	0.500	0.470	0.559	0.490	0.513	0.488	0.525	0.478
Decelerate	0.255	0.242	0.284	0.241	0.256	0.246	0.266	0.245	0.263
Adjust steering left	0.351	0.386	0.324	0.356	0.369	0.366	0.367	0.375	0.355
Maintain steering	0.504	0.438	0.503	0.470	0.464	0.460	0.476	0.456	0.478
Adjust steering right	0.145	0.175	0.172	0.175	0.166	0.174	0.156	0.169	0.167
Turn left	0.093	0.095	0.118	0.076	0.105	0.093	0.111	0.088	0.113
No turn	0.896	0.864	0.834	0.908	0.855	0.875	0.847	0.873	0.855
Turn right	0.012	0.041	0.047	0.016	0.041	0.032	0.041	0.039	0.031
Total cases	345	906	296	315	1232	1016	531	877	670

- Route 1: Sunoco Gas Station→ 15180 Washington St→ UVA Health Haymarket → Budget Truck Rental → Historic District → 15180 Washington St → Sunoco Gas Station → 7502 Iron Bar Ln
- Route 2: 7502 Iron Bar Ln → Budget Truck Rental → 15180 Washington St → 7502 Iron Bar Ln → UVA Health Haymarket → Historic District → Sunoco Gas Station

Each participate are asked to drive through both route, at different daytimes and weather conditions. In general, we conducted 28 daytime and 20 evening driving sessions.

For the raw collected data, there are 3 main categories: front-view video recordings, vehicle dynamic information and gaze movements of driver. For front-view video recordings, we mount a camera on top of the vehicle that recording the front-view scenario during driving to simulate drivers' perceived environment. We totally collect 2,537,067 images from all recordings. For vehicle dynamic information, we utilize the CAN bus interface to collect vehicle motion and control data (acceleration, velocity, steering angle, throttle position, etc.) at 50 Hz sampling rate from the vehicle's Electronic Control Unit (ECU), while GPS devices record vehicle trajectory information. For drivers' gaze movements, we employ a dashboard-mounted SmartEye Pro v5.0 system featuring dual infrared cameras to track driver face and eye movements. The system performs head tracking by estimating the 3D pose of the subject's head and eye tracking by recording gaze origin and direction vectors for both eyes. These are combined to generate consensus gaze directions, which are then projected onto the visual scene at 50 Hz to determine where drivers focus their attention.

Dataset preprocess. Due to unavoidable gaze projection failures under natural driving conditions (e.g., blinking or head movements), a portion of frames does not contain valid gaze data. As summarized in Table 5, 31.5% of daytime frames and 32.0% of evening frames were filtered out. In addition, due to different weather conditions, there are some camera misfocus which leads to blurriness of the image, which also have been filtered out. The remaining data were manually curated into 1,500 short video clips (3 s each, 75 frames), resulting in a final dataset of approximately 110K frames used for downstream analysis. For the selected data, it covers drivers across multiple demographic dimensions, including gender, age groups (≤25 and >25), and driving experience levels (≤5 years and >5 years). As summarized in Table 6, the distribution of driving behaviors is relatively balanced across these groups, with no single demographic or driving style dominating the dataset. This diversity enables reliable analysis of human driving behavior across different population segments. In addition to demographic diversity, our selected dataset spans a variety of driving scenarios, including freeway, ramp, and local road segments, as well as conditions such as dusk, heavy rain, and heavy traffic. Table 7 summarizes the distribution of scenarios and the corresponding driving behavior frequencies across these conditions.

4.2. Dataset annotation

MLLM annotation. To bridge the gap between raw sensory data and semantic understanding, we implement a structured annotation process covering six key components: *driver*, *environment*, *host vehicle*, *traffic dynamics*, *navigation*, and *human gaze attention*. Directly available data (e.g., driver age) populate relevant template fields, while additional details are derived through traffic expert annotations. For environment information, we input sequential scene images (captured at 1-s intervals) into GPT-4o to generate textual descriptions. To ensure accuracy and minimize hallucinations, expert-annotated environment descriptions are provided as In-Context Learning examples, thereby standardizing GPT-4o outputs. Similarly, gaze information is summarized by overlaying driver attention maps onto scene images, highlighting driver-focused areas.

Table 7
Scenario diversity and behavior distribution across different driving conditions.

Statistic	Freeway	Ramp	Local	Day	Dusk	Heavy Rain	Heavy Traffic
Sample count	717	204	626	822	725	249	82
Avg. words per prompt	330.5	321.4	325.2	327.4	326.4	327.3	351.0
Speed control							
Maintain speed (%)	62.8	37.7	40.4	51.9	48.7	41.6	36.6
Decelerate (%)	16.2	37.3	31.8	24.3	26.3	26.9	51.2
Accelerate (%)	21.1	25.0	27.8	23.7	25.0	31.5	12.2
Turning decision							
No turn (%)	99.3	96.6	68.7	87.5	85.5	80.5	85.5
Turn left (%)	0.6	1.0	23.5	9.7	10.1	14.1	8.4
Turn right (%)	0.1	2.5	7.8	2.8	4.4	5.4	4.8
Steering decision							
Maintain steering (%)	28.2	52.9	65.5	48.7	44.1	38.9	58.5
Adjust left (%)	44.6	36.3	27.6	35.4	38.1	37.6	31.7
Adjust right (%)	27.2	10.8	6.9	15.9	17.8	23.5	9.8

Expert annotation. For detailed and diverse traffic dynamics annotations, we rely on human annotators using structured templates. Sixteen traffic engineers annotate video segments, detailing lane topology, relative states of surrounding vehicles (distance, velocity changes, motion trajectories), and other dynamics. These annotations are systematically transformed into coherent textual descriptions, ensuring consistency and completeness across varied scenarios. More specifically, we discretize longitudinal control into three mutually exclusive classes: Accelerate, Decelerate, No-meaningful-change. To focus labeling on moments with interpretable intent (rather than noise), we first automatically extract candidate timestamps where the vehicle speed changes by $>10\%$ within 1s, and the brake or throttle input changes concurrently. For each candidate event, we generate a 10-s clip centered on the event and provide annotators the synchronized video plus vehicle-dynamics traces; annotators then assign one of the three classes based on whether the behavior reflects intentional acceleration, intentional deceleration, or normal fluctuation (no meaningful change). Besides, we use a dedicated annotation platform supporting both labeling and review. Annotators follow a written guideline document specifying criteria for each category with representative examples. Each expert performs (i) initial annotations and (ii) reviews annotations made by others; after an annotation is completed, it is randomly assigned to a different expert for review and correction, providing a structured cross-check to promote consistency. Fig. 4 illustrates the annotation interface used in our paper, which presents key segments of driving scene clips together with relevant driving signals (e.g., vehicle state and environment-related data) to support informed and accurate annotation decisions. To ensure annotation consistency and structural uniformity, the platform enforces a closed-set annotation protocol, where annotators assign labels by selecting from predefined categories rather than providing free-form inputs.

5. Benchmarks and experiments summary

5.1. Evaluated models and experimental setup

We evaluate a range of widely used commercial and open-source LLMs that are commonly employed in autonomous driving and LLM-based driving agent research to assess model performance on our human-centric driving dataset (Yang et al., 2023a). The commercial models are grouped into three series: the **GPT series** (GPT-4o, GPT-4o mini, and GPT-o3 mini) from OpenAI (Hurst et al., 2024), the **Gemini series** (Gemini 2.0 Flash and Gemini 2.5 Flash) from Google (Team et al., 2023), and the **Claude series** (Claude 3.5 Haiku and Claude 3.7 Sonnet) from Anthropic (Anthropic, 2024). The open-source models include the **Llama series** (Llama 4 Scout, Llama 4 Maverick, and Llama 3) developed by Meta (Touvron et al., 2023; Singh, 2025) and **Qwen series** (Qwen-Max, Qwen-VL-Max, Qwen-VL-Plus) from Alibaba (Bai et al., 2023). Each model is evaluated using identical inputs across all cases, comprising a uniform text prompt and its paired driving scene image. All selected LLMs were accessed through their official APIs using the most recent publicly available endpoints as of April 2025. Unless otherwise specified, default decoding parameters were applied across all experiments, with temperature and top-p set to 1.0. Fig. 5 illustrates a prompt examples from the HermesSense dataset. These prompts serve as structured textual inputs to the LLM-based agent, encoding multimodal driving context to facilitate reasoning and behavior prediction.

5.2. Evaluation of decision alignment with human drivers

We evaluate each LLM's performance by comparing two input modalities, using text input alone and using textual data with image inputs. We randomly sampled 1,500 instances from the full dataset for evaluation. The results for LLMs' decision alignment are shown at Table 8.

Comparison of decision alignment for different LLMs . We compare the decision alignment performance of various LLMs across three driving decision tasks. Overall, no single model consistently outperforms all others across all tasks, while Claude 3.7 Sonnet demonstrates strong performance in general, achieving the highest F1 scores on the speed and steering tasks, and the second-highest F1 score on the turning task. In the speed task, Llama 4 Maverick and Llama 4 Scout

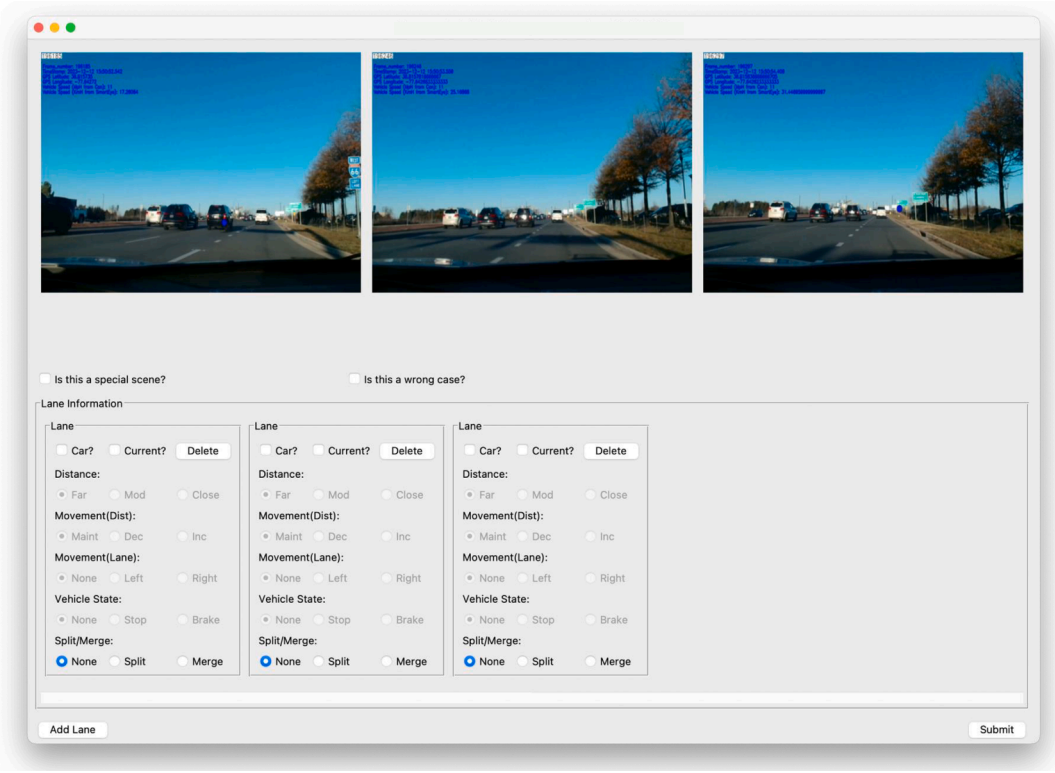


Fig. 4. Annotation platform illustrating the driving scene clips and vehicle-dynamics traces used for driving behavior labeling.

Table 8
Decision alignment results of LLMs across three driving decision tasks. Each cell presents paired results in the format: text-only input / text-plus-image input. *better* values indicate performance improvements compare to text-only input, while worse values indicate performance decreases. The model with the best performance is shown in bold. Llama3 70B, Qwen-textual model only supports textual inputs.

Models	Speed			Turning			Steering		
	F1 (%) ↑	Acc. (%) ↑	$\mathbb{D}_{ali}$ (×100) ↓	F1 (%) ↑	Acc. (%) ↑	$\mathbb{D}_{ali}$ (×100) ↓	F1 (%) ↑	Acc. (%) ↑	$\mathbb{D}_{ali}$ (×100) ↓
GPT-4o	35.8 / *36.0*	48.5 / *49.4*	9.8 / 11.2	65.2 / *67.3*	82.8 / *84.2*	2.0 / *1.7*	55.0 / *55.6*	58.0 / *58.9*	2.6 / 4.0
GPT-4o mini	37.0 / *38.3*	52.6 / *54.3*	12.5 / 13.4	56.7 / 55.6	74.7 / 73.7	5.1 / 5.2	45.9 / 43.0	46.2 / 44.1	1.1 / 1.8
GPT-o3 mini	43.7 / 38.2	**52.7** / *54.4*	4.1 / 13.8	68.0 / 53.8	84.7 / 71.9	1.2 / 5.9	47.9 / 43.1	50.0 / 44.0	2.5 / *1.5*
Claude 3.7 Sonnet	**48.5** / 48.5	50.0 / *50.8*	2.8 / *1.5*	74.4 / 71.6	88.0 / 86.4	0.9 / 1.3	**58.0** / **57.6**	**60.0** / ***60.1***	1.6 / 2.0
Claude 3.5 Haiku	38.3 / *39.7*	51.9 / *54.1*	12.2 / 12.8	51.9 / *54.6*	68.0 / *71.0*	7.9 / *6.6*	50.5 / 40.6	56.5 / 51.8	7.5 / 14.6
Gemini 2.5 Flash	44.8 / *46.4*	46.2 / *47.7*	4.5 / *3.8*	70.4 / 69.3	86.8 / 85.9	0.7 / 1.1	48.8 / *49.9*	53.4 / *53.8*	4.4 / *3.7*
Gemini 2.0 Flash	42.3 / 40.0	49.5 / 47.4	3.9 / 6.1	67.5 / 65.4	84.9 / 82.5	1.5 / 2.3	52.9 / *54.6*	55.1 / *56.6*	2.0 / ***1.1***
Llama 4 Maverick	47.1 / ***49.0***	49.3 / *53.5*	**1.0** / ***0.3***	61.6 / *62.4*	78.0 / *79.8*	3.6 / *2.9*	52.8 / 49.8	53.9 / 51.5	**0.5** / 1.2
Llama 4 Scout	47.0 / *48.9*	52.2 / ***54.9***	1.3 / 1.9	**79.1** / **77.0**	**92.4** / **91.7**	**0.1** / **0.1**	42.3 / *43.2*	46.9 / 45.8	5.3 / *4.2*
Llama 3 70B	41.3 / –	49.3 / –	7.2 / –	65.1 / –	81.8 / –	2.5 /–	49.5 /–	52.5 /–	2.5 / –
Qwen-vl-max	44.1 / 43.2	46.9 / *47.6*	3.2 / *2.5*	72.0 / 65.5	89.8 / 89.3	0.1 / 0.1	51.3 / 49.1	53.7 / 52.7	1.9 / 2.9
Qwen-vl-plus	29.8 / *31.9*	52.0 / 49.6	20.9 / *16.0*	53.5 / 48.6	87.3 / 87.3	1.4 / 2.7	30.4 / *34.2*	42.7 / 40.9	14.6 / *11.7*
Qwen-textual	42.8 / –	46.9 / –	3.6 / –	52.9 / –	69.9 / –	6.9 /–	50.7 /–	50.6 /–	1.2 / –

are comparable to Claude 3.7 Sonnet and exhibit the best distributional alignment scores (1.0 and 1.3, respectively). For the turning task, Llama 4 Scout shows notable improvements over other models, with an F1 score of 0.791 (+19.7% relative to task average). In the steering task, except for Claude 3.7 Sonnet (highest F1 score and accuracy, 0.58 and 0.60, respectively), Llama 4 Maverick attains the best distributional alignment score (0.5). Overall, models with larger parameter sizes (GPT-4o, Claude 3.7 Sonnet, Gemini 2.5 Flash, and Llama 4 Maverick) tend to outperform those with fewer parameters across most evaluation metrics.

Table 8 presents a comparative analysis of how incorporating driving scene images as additional input affects the decision alignment across LLMs. The addition of visual information does not uniformly enhance model performance across tasks. For instance, Llama 4 Maverick demonstrates improved performance metrics on the speed task when provided with visual input, yet exhibits performance degradation on the steering task under the same multimodal conditions. Similarly, the Gemini family of models displays

Example of organized prompt for #846

Driving Scenario Description

[Driver demographic information]:
Assume you are a 26-year-old Hispanic/Latino male driver with 2 years of driving experience, driving approximately 30 miles per week. You will be provided with descriptions of real-world driving scenarios. Based on these details, simulate the driver's decision-making and behavior.

[Environment information]:
The driving environment is at an intersection with overcast weather conditions and diffuse daylight. The road is flat with no visible inclines or curves. Lane markings are present, dividing the lanes clearly. The road surface appears to be in fair condition, with some visible debris near the edges. Traffic signals are functioning, displaying green lights, and there are poles on either side of the intersection. There is a sign prohibiting left turns on the side. Additional signage can be seen further ahead providing direction. No visible construction zones, barriers, or pedestrian crossings are present. There is moderate traffic density, with vehicles in adjacent lanes moving through the intersection. The traffic flow appears steady, with no obstructions evident.

[Vehicle dynamics & Driver behavior information]:
The vehicle is maintaining a steady speed of 47 kilometers per hour. The driver's throttle input over 3 seconds was ['low', 'low', 'high'], and the brake input was ['none', 'none', 'none'].

[Traffic dynamics information]:
There are 3 lanes on the road. The ego vehicle is driving in the middle lane.
In the leftmost lane, there is a vehicle at Close distance, the gap is widening. In the middle lane, there is no vehicle. In the rightmost lane, there is no vehicle.

[Navigation information]:
Navigation suggests using the left 1 lanes to take the left exit. The exit is ahead at a moderate distance.

[Gaze information]:
In the gaze-highlighted regions of both images, a blue car is visible in the center lane of a multi-lane highway surrounded by a concrete barrier on the right. The sky above is clear with scattered clouds, and there are faint outlines of distant vehicles further down the road. The road is marked with white dashed lines, and a lamppost is visible in the background to the right.

Task Instruction

Based on the provided driving scenario, select the most likely driver action at each step. After selecting, provide a reason for your choice based on road conditions, traffic flow, and driving safety considerations.
Step 1: Speed Adjustment
- Maintain speed - Accelerate - Decelerate
Step 2: Turning Decision (This decision is made specifically at intersections and does not occur simultaneously with a steering change.)
- Turn left - Turn right - No turn
Step 3: Steering Change Decision (Including lane change maneuvers and evasive steering.)
- Adjust the steering to left - Adjust the steering to right - Maintain the steering

---

The ground truth is: *<Accelerate>, <No turn>, <Adjust the steering to left>*

Fig. 5. A prompt example for #846 instance.

high sensitivity to visual inputs, with significant performance fluctuations across different driving tasks when visual information is incorporated (for Gemini 2.5 Flash, increase in Speed and steering tasks while decrease in turning task).

Decision alignment across diverse driving contexts. Following prior work highlighting the impact of contextual diversity on decision-making in driving scenarios (Zeng et al., 2019; Janai et al., 2020), a fine-grained evaluation of LLM decisions using our text input is conducted to assess their alignment with human drivers under varying environmental conditions. We partitioned our dataset based on three critical contextual factors: 1) temporal conditions (daytime / dusk or night), 2) intersection presence (Yes / No), and 3) road type (freeway, local road, and ramp) (Ding et al., 2019; Deo and Trivedi, 2018). For each subset, we computed both the F1 score and distribution alignment score $\mathbb{D}_{ali}$ to quantify the decision alignment under different scenarios. In general, Claude 3.7 Sonnet demonstrates higher F1 score across most scenarios, especially in the *not at intersection* condition, where it outperforms all other models across all three tasks. In dusk/night and ramp conditions, Llama 4 Maverick exhibits highest correctness, particularly in the steering and turning tasks. From the perspective of distributional alignment, the Llama family and Claude 3.7 Sonnet demonstrate strong performance across most scenarios, achieving $\mathbb{D}_{ali}$ of 0.035 and 0.023, re-

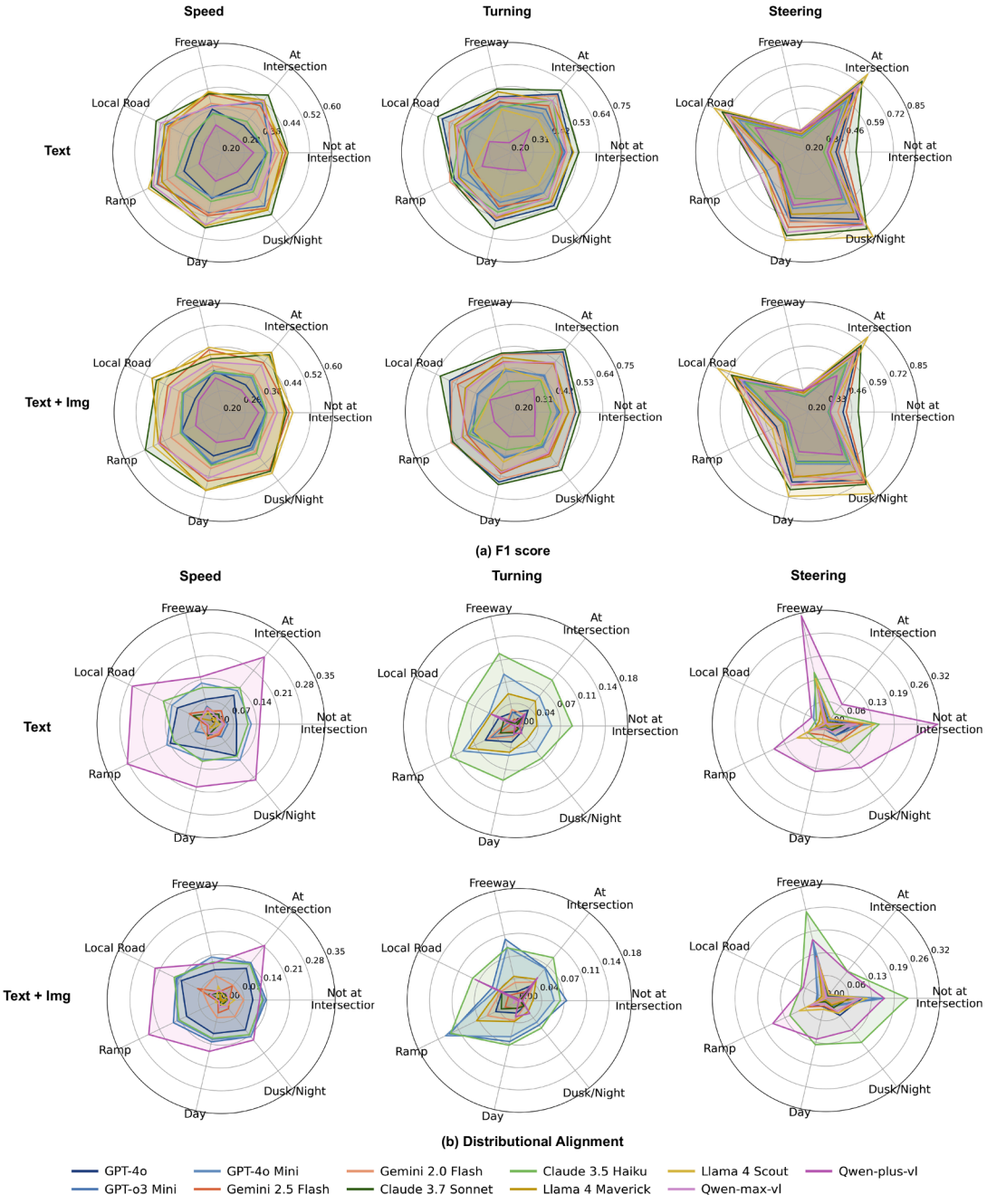


Fig. 6. LLMs' decision alignment across various scenarios. We evaluate model performance on speed control, turning, and steering decisions using F1 score and $\mathbb{D}_{\text{ali}}$ across varying conditions, including time of day, location type, and intersection presence. Higher F1 scores indicate better behavioral alignment, while lower $\mathbb{D}_{\text{ali}}$ reflect greater distributional similarity.

spectively, compared to the overall average of 0.053. Notably, in the steering task, we observe that the majority of LLMs exhibit improved alignment with human decision distributions under the *at intersection* and *local road* conditions.

Effect of visual input on alignment. We further assess the influence of image input by comparing model alignment under text-only and text + image conditions, as presented in the bottom panel of Fig. 6. Performance varies across model families: the Qwen series benefits from visual input, with a mean JSD reduction of 0.006, indicating improved alignment. In contrast, GPT series models show a degradation in distributional alignment, with an average JSD increase of 0.0114. Claude 3.7 Sonnet remains largely invariant to

Table 9
Feature attribution results across F1 and $\mathbb{D}_{ali}$ metrics for all driving decision tasks. Parentheses indicate the rank of each component's contribution. R_h denotes human-perceived importance rankings, $\mathbb{D}_{ali}$ measures divergence alignment, $\mathbb{A}_{sim}$ reflects attribution similarity and R_{avg} denotes the average rank of each component's contribution. Results are reported for four models: Claude 3.7 Sonnet, GPT-4o mini, Gemini 2.0 Flash, and Llama 4 Maverick.

Task	Component	Claude			Llama			GPT			Gemini			R_h
		F1↑	$\mathbb{D}_{ali}$↓	R_{avg}	F1↑	$\mathbb{D}_{ali}$↓	R_{avg}	F1↑	$\mathbb{D}_{ali}$↓	R_{avg}	F1↑	$\mathbb{D}_{ali}$↓	R_{avg}	
Speed	Driver Info	0.04 (5)	−0.06 (6)	5.5	−0.01 (6)	0.00 (6)	6.0	0.02 (6)	−0.05 (6)	6.0	−0.01 (6)	−0.02 (6)	6.0	6
	Env. Info	0.07 (3)	−0.16 (3)	3.0	0.11 (2)	−0.22 (2)	2.0	0.07 (3)	**−0.20 (1)**	2.0	0.07 (3)	−0.18 (3)	3.0	2
	Vehicle Dynamics	**0.16 (1)**	**−0.25 (1)**	1.0	**0.16 (1)**	**−0.29 (1)**	1.0	0.08 (2)	−0.19 (2)	2.0	**0.11 (1)**	**−0.27 (1)**	1.0	5
	Traffic Dynamics	0.12 (2)	−0.18 (2)	2.0	0.10 (3)	−0.19 (3)	3.0	**0.11 (1)**	−0.17 (4)	2.5	0.08 (2)	−0.18 (2)	2.0	1
	Navi. Info	0.06 (4)	−0.15 (4)	4.0	0.02 (5)	−0.09 (5)	5.0	0.03 (5)	−0.09 (5)	5.0	0.04 (4)	−0.11 (5)	4.5	4
	Attention Info	0.03 (6)	−0.14 (5)	5.5	0.04 (4)	−0.17 (4)	4.0	0.04 (4)	−0.16 (3)	3.5	0.04 (5)	−0.17 (4)	4.5	3
	Total Attribution	0.48	−0.94	–	0.43	−0.95	–	0.35	−0.85	–	0.32	−0.93	–	–
	$\mathbb{A}_{sim}$	0.20	0.33	0.27	0.33	0.33	0.33	0.60	0.33	0.47	0.47	0.33	0.40	–
Turn	Driver Info	0.04 (6)	−0.06 (6)	6.0	0.00 (6)	0.00 (6)	6.0	0.03 (6)	−0.06 (6)	6.0	0.01 (6)	−0.04 (6)	6.0	6
	Env. Info	0.10 (2)	−0.20 (2)	2.0	0.08 (2)	−0.20 (2)	2.0	0.10 (2)	−0.20 (2)	2.0	0.08 (2)	−0.20 (2)	2.0	3
	Vehicle Dynamics	0.07 (3)	−0.16 (4)	3.5	0.06 (3)	−0.18 (4)	3.5	0.05 (3)	−0.16 (4)	3.5	0.05 (4)	−0.18 (3)	3.5	5
	Traffic Dynamics	0.07 (4)	−0.15 (5)	4.5	0.04 (5)	−0.17 (5)	5.0	0.08 (4)	−0.16 (5)	4.5	0.06 (3)	−0.17 (5)	4.0	2
	Navi. Info	**0.51 (1)**	**−0.26 (1)**	1.0	**0.52 (1)**	**−0.26 (1)**	1.0	**0.42 (1)**	**−0.22 (1)**	1.0	**0.35 (1)**	**−0.24 (1)**	1.0	1
	Attention Info	0.05 (5)	−0.17 (3)	4.0	0.05 (4)	−0.19 (3)	3.5	0.04 (5)	−0.17 (3)	4.0	0.04 (5)	−0.17 (4)	4.5	4
	Total Attribution	0.83	−1.00	–	0.75	−1.00	–	0.71	−0.98	–	0.59	−1.00	–	–
	$\mathbb{A}_{sim}$	0.73	0.60	0.67	0.47	0.60	0.54	0.73	0.73	0.73	0.73	0.60	0.67	–
Steer	Driver Info	0.05 (4)	−0.06 (6)	5.0	−0.01 (6)	0.00 (6)	6.0	0.02 (6)	−0.07 (6)	6.0	0.00 (6)	−0.03 (6)	6.0	6
	Env. Info	0.04 (5)	−0.19 (2)	3.5	0.03 (4)	−0.20 (3)	3.5	0.06 (3)	−0.20 (2)	2.5	0.04 (5)	−0.21 (2)	3.5	3
	Vehicle Dynamics	0.04 (6)	−0.16 (5)	5.5	0.05 (2)	−0.19 (4)	3.0	0.05 (4)	−0.16 (4)	4.0	0.04 (4)	−0.18 (4)	4.0	5
	Traffic Dynamics	0.17 (2)	**−0.22 (1)**	1.5	**0.13 (1)**	**−0.24 (1)**	1.0	0.09 (2)	**−0.20 (1)**	1.5	0.11 (2)	**−0.23 (1)**	1.5	1
	Navi. Info	**0.18 (1)**	−0.18 (4)	2.5	0.03 (5)	−0.15 (5)	5.0	**0.12 (1)**	−0.16 (5)	3.0	**0.12 (1)**	−0.15 (5)	3.0	2
	Attention Info	0.07 (3)	−0.19 (2)	2.5	0.05 (3)	−0.21 (2)	2.5	0.05 (5)	−0.19 (3)	4.0	0.05 (3)	−0.20 (3)	3.0	4
	Total Attribution	0.55	−0.99	–	0.38	−1.00	–	0.40	−0.97	–	0.37	−1.00	–	–
	$\mathbb{A}_{sim}$	0.47	0.60	0.54	0.47	0.47	0.47	0.73	0.60	0.67	0.60	0.60	0.60	–

image input, exhibiting minimal F1 change (e.g., +0.0001 for speed, -0.0011 for steering), suggesting a robust yet conservative integration of visual cues.

5.3. Attribution-based human-likeness evaluation

Following the method described in Section 3.1, we evaluate the alignment between model-derived feature attributions and the information prioritization by human drivers. We randomly sampled 500 instances from our complete dataset for this analysis. To establish human-prioritized reference rankings, we implemented a manual annotation protocol in which 20 domain experts independently annotated ten sets of feature importance rankings for each decision class within each driving task. The feature attribution results of the LLM agents, together with their similarity to these human-annotated references, are summarized in Table 9.

Claude and GPT benefit from driver personas. We observe that different models exhibit varying sensitivity to specific types of information. Claude 3.7 Sonnet and GPT-4o mini show improved performance when provided with driver demographic inputs, with such information consistently enhancing their performance across all evaluation metrics. For example, providing driver demographic description increases Claude's performance in the speed task by 8.3%. In contrast, Llama 4 Scout shows limited reliance on driver demographic inputs, with minimal performance changes. Moreover, we find that this pattern holds consistently across different tasks, when driver-related inputs receive attribution in one task, they tend to exert a similarly positive influence on model performance in other tasks as well. The similar pattern can be seen from genmini 2.0 Flash and Llama 4 Scout.

LLMs' attribution align with human-perceived priorities in turning and steering tasks. The attribution similarity $\mathbb{A}_{sim}$ is used to evaluate the similarity between the importance rankings of different information types derived from model data attribution and those based on human-perceived information importance (see Eq. (6)). From Table 9, we observe that in the turning and steering tasks, the models achieve average $\mathbb{A}_{sim}$ of 0.649 and 0.568, respectively, indicating a high degree of alignment with human information prioritization. In contrast, for the speed task, the models exhibit a lower average $\mathbb{A}_{sim}$ of 0.365. Specifically, the models place greater emphasis on vehicle dynamics information compared to human drivers.

The type of information that drives alignment varies across tasks and decision types. Within each task, all four LLMs produce similar attribution patterns, as indicated by the consistency in their feature rankings. However, the dominant information sources vary substantially by task. In the speed control task, vehicle dynamics and traffic dynamics jointly account for an average of 59.0% of the

Table 10
Feature attribution analysis of F1 scores for label-level predictions within the driving decision tasks *Speed*, *Turn*, and *Steer*. Attribution values quantify the contribution of each input component to the model's ability to correctly predict specific control decisions (e.g., *Accelerate*, *Turn Left*, *Maintain*). Parentheses denote the rank of each component's influence for the given label. R_{avg} denotes the average rank across the three labels of each task, while R_h indicates human-assessed importance rankings. The $\mathbb{A}_{\text{sim}}$ metric captures attribution similarity between model reasoning and human expectations. All results are obtained from predictions made by the Claude 3.7 Sonnet model.

Component	Speed				Turn				Steer			
	Accelerate	Maintain	Decelerate	$\mathbf{R_{avg}/R_h}$	Turn Left	No Turn	Turn Right	$\mathbf{R_{avg}/R_h}$	Adjust Left	Maintain	Adjust Right	$\mathbf{R_{avg}/R_h}$
Driver Info	0.01 (4)	0.05 (6)	0.01 (6)	5.33 / 6	-0.01 (5)	0.11 (6)	0.00 (6)	5.67 / 6	0.00 (5)	0.14 (3)	0.00 (5)	4.33 / 6
Env. Info	0.05 (3)	0.10 (3)	0.09 (3)	3.00 / 2	0.08 (2)	0.15 (2)	0.08 (2)	2.00 / 3	0.01 (4)	0.11 (5)	0.02 (4)	4.33 / 3
Vehicle Dyn.	**0.28 (1)**	0.11 (2)	0.10 (2)	1.67 / 5	0.04 (3)	0.15 (2)	0.01 (4)	3.00 / 5	0.00 (5)	0.15 (2)	-0.01 (6)	4.33 / 5
Traffic Dyn.	0.11 (2)	**0.12 (1)**	**0.13** (1)	1.33 / 1	0.03 (4)	0.15 (2)	0.01 (4)	3.33 / 2	0.19 (2)	**0.22 (1)**	0.10 (2)	1.67 / 1
Navi. Info	0.00 (5)	0.09 (4)	0.08 (4)	4.33 / 4	**0.63 (1)**	**0.22 (1)**	**0.69 (1)**	1.00 / 1	**0.21 (1)**	-0.01 (6)	**0.34 (1)**	2.67 / 2
Attention Info	0.00 (5)	0.07 (5)	0.04 (5)	5.00 / 3	-0.02 (6)	0.14 (5)	0.02 (3)	4.67 / 4	0.03 (3)	0.14 (3)	0.03 (3)	3.00 / 4
Total Attribution	0.45	0.54	0.46	–	0.76	0.92	0.81	–	0.43	0.75	0.47	–
$\mathbb{A}_{\text{sim}}$	0.07	0.47	0.47	–	0.47	0.73	0.73	–	0.60	-0.07	0.60	–

total contribution, whereas navigation information dominates the turning task with an average contribution of 62.5%. In the steering task, traffic dynamics and navigation information together contribute 61.8%, highlighting the task-dependent nature of information utilization. Table 10 presents the fine-grained, decision-level attribution analysis, revealing distinct patterns of information utilization across different control actions. Specifically, driver attention exhibits a comparatively higher average contribution (0.12) for decisions involving maintaining current behavior (maintaining speed, steering, or no turn), compared to other decision classes (0.02). This suggests that driver gaze provides more consistent or discriminative signals during steady-state driving, which may help LLM agents better capture human driving patterns and reduce mispredicted maneuvers such as unintended lane changes or turns. Besides, despite the varying dominance of features across tasks and decisions, no input feature receives significant negative attribution across models, indicating that richer contextual information does not degrade performance and is generally safe for guiding model decisions.

5.4. How good are LLM agents at mimicking various types of drivers

To assess the human-likeness of LLM agents across diverse driver types, we group drivers based on demographic and behavioral attributes. Fig. 7 illustrates the statistical distributions of throttle input, vehicle speed, and braking behavior exhibited by different driver types across multiple trips. For every driver type, we compute behavioral similarity scores (as defined in Section 3.1). Fig. 8 shows the long-term human-likeness of all LLM agents.

LLM mimic aggressive driver better. Overall, we observe that LLMs tend to better simulate drivers who are female, under the age of 25, have a history of accidents, have more than five years of driving experience, and exhibit a more aggressive driving style. Among all models, Claude 3.7 Sonnet demonstrates the strongest overall ability to simulate most driver types. However, for drivers aged 25 or younger, those with accident histories, and those exhibiting more aggressive driving styles, Llama 4 Scout and Gemini 2.0 show better human-likeness.

Relationship between sequence length and statistical human-likeness. The Statistical Human-likeness metric evaluates temporal consistency and pattern similarity over fixed-length sequences rather than long-term future actions. Selecting the sequence length L is a trade-off: short sequences fail to capture stable behavior (increasing variance in $\mathbb{B}_{\text{sim}}$B), while excessively long sequences introduce noise from context shifts and increase computational overhead. In our experiments, the action sequence length L is set to 50 consecutive decision steps, corresponding to approximately a 5-minute driving segment, during which the navigation goal and traffic context remain largely consistent. This setting allows the metric to effectively capture human-like behavioral inertia and decision patterns while avoiding noise accumulation caused by scene transitions. To further clarify the impact of sequence length on the experimental results, we additionally conduct an ablation study with sequence lengths ranging from $L = 5$ to $L = 65$ with a step size of 2. Fig. 9 further illustrates the trend of $\mathbb{B}$sim with respect to the sequence length L for six representative models across different driving styles. Tables X report the detailed $\mathbb{B}_{\text{sim}}$ values and their variances across the five driver demographic groups. As shown in both Table X and Figure X, for most models $\mathbb{B}_{\text{sim}}$ exhibits relatively large fluctuations when L is small, mainly due to higher variance under short action sequences, and then gradually stabilizes as L increases.In particular, when $L \geq 40$, the variance of $\mathbb{B}_{\text{sim}}$ decreases by 33.5% on average.

Modality effects on driver-type simulation. Fig. 10 illustrates the impact of incorporating images as input on LLMs' ability to simulate different types of drivers. With the addition of image input, Claude 3.7 shows an improvement in simulating drivers with an accident history and those exhibiting aggressive driving styles, while GPT-4o demonstrates enhanced performance in simulating defensive and female drivers. In contrast, incorporating image input leads to a decrease in the human-likeness of Gemini 2.0 and GPT-o3 Mini across different driver types.

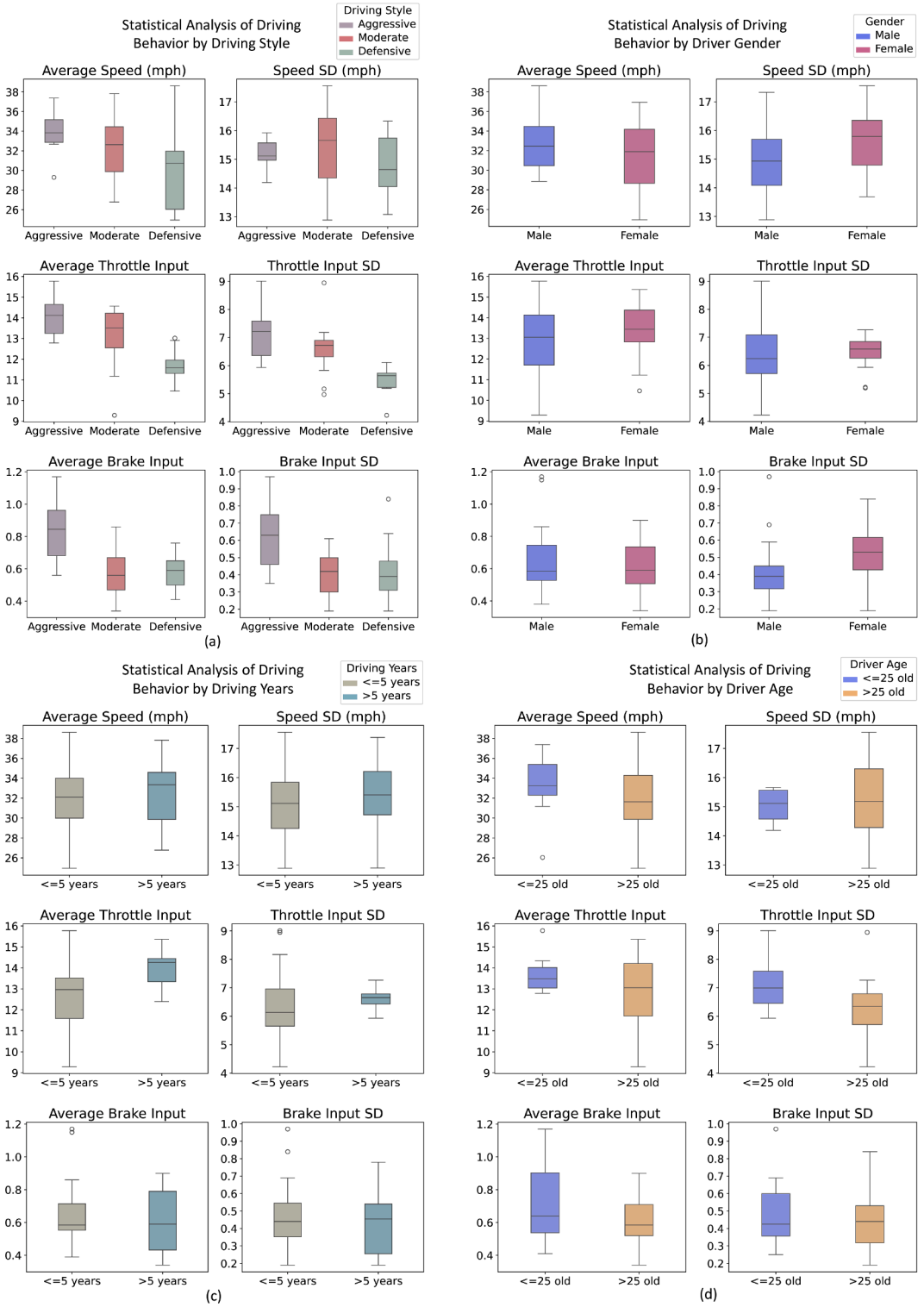


Fig. 7. Statistical analysis of driving behavior across different groups. (a) By driving style: aggressive, moderate, and defensive. (b) By gender: female vs. male drivers. (c) By driving experience: less vs. more experienced drivers. (d) By driver age: younger vs. older drivers. Each subfigure summarizes the average and variability of brake input, throttle input, and speed.

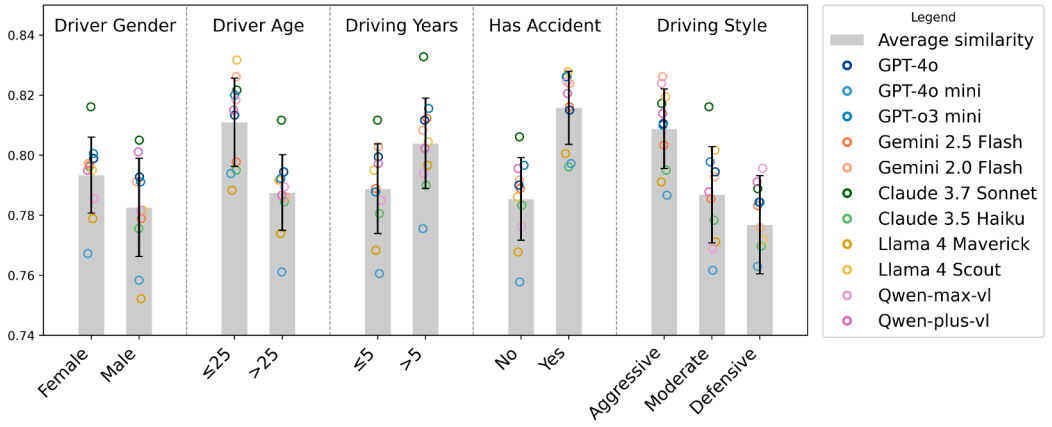


Fig. 8. Results of the human-likeness of LLMs across various driver types. Overall statistical human-likeness of all LLM agents across different driver profiles. Each circle represents an individual LLM agent's human-likeness score.

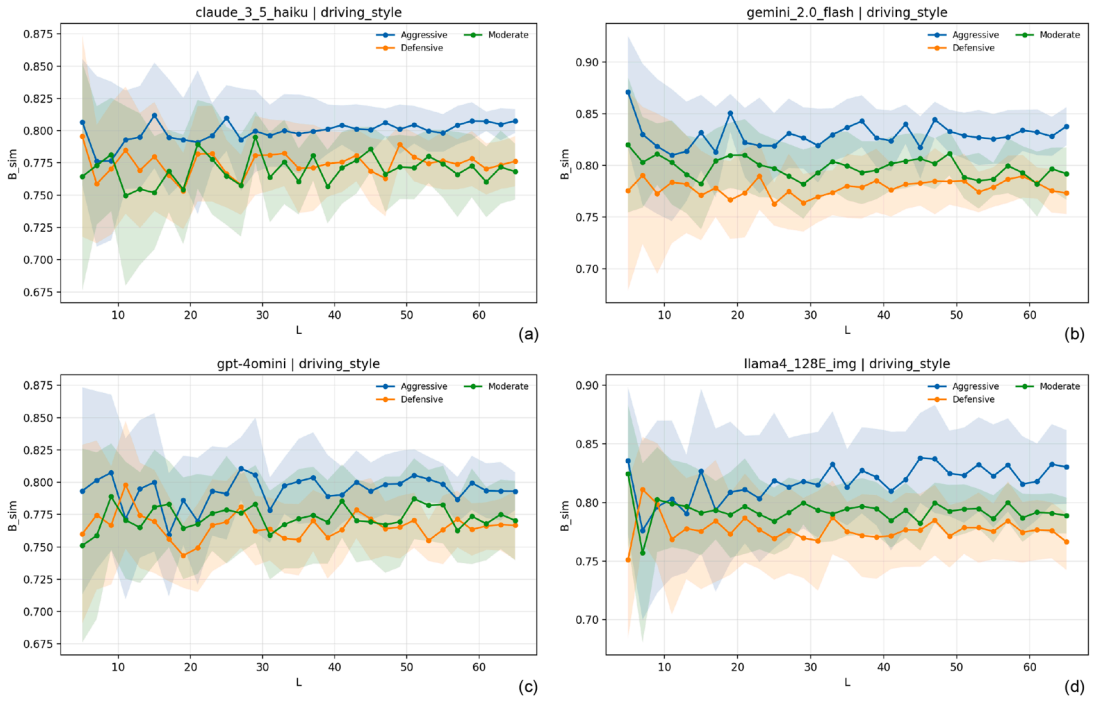


Fig. 9. Sensitivity of the Statistical Human-likeness $\mathbb{B}\text{sim}$ to the action sequence length L across different driving styles. The figure reports results for four representative models: (a) Claude-3.5-Haiku, (b) Gemini-2.0-Flash, (c) GPT-4o-mini, and (d) LLaMA-4-128E-Img. For each model, $\mathbb{B}\text{sim}$ is evaluated under aggressive, defensive, and moderate driving styles. Solid lines denote the mean $\mathbb{B}\text{sim}$ across driver demographic groups, while shaded regions indicate the corresponding variance.

5.5. Analysis of performance degradation induced by image input

Although image inputs are often assumed to provide richer contextual detail than text alone, and multimodal inputs have the potential to enhance model understanding in complex tasks, our experimental results across both the short-term alignment evaluation (Section 5.2) and long-term human-likeness evaluation (Section 5.4) consistently show that incorporating visual information does not universally improve model performance. On certain models, the addition of image inputs leads to measurable degradation. This phenomenon warrants closer examination. Below, we identify two primary contributing factors:

Data modality interference in MLLMs. Recent studies show that MLLMs often struggle to distinguish task-relevant from irrelevant modality inputs and cross-modal fusion may introduce unintended interference, thereby degrading unimodal performance (Cai et al., 2025). In particular, such interference may stem from challenges inherent during the vision-language alignment process, including

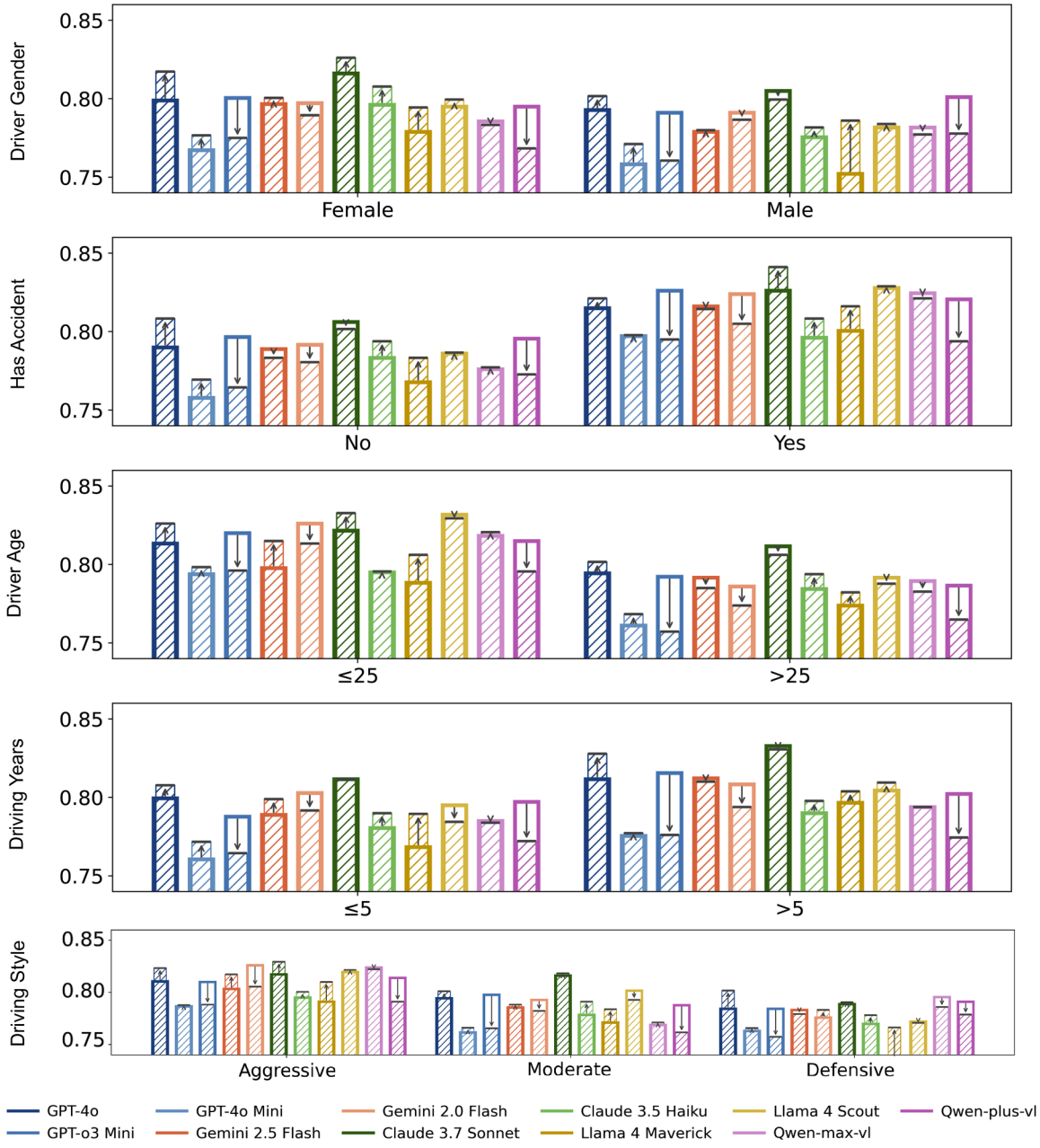


Fig. 10. Statistical human-likeness change for the LLMs per driver type when augmented with image input.

catastrophic forgetting, knowledge conflicts, and spurious correlations (Chen et al., 2024a). In our task, driving scenes typically involve a larger amount of dynamic entities and complex interactions compared to standard VQA tasks like classification. However, much of this visual information does not directly inform the decision task, for example, roadside advertisements or parked vehicles in the opposite lane may be visually salient yet irrelevant to the current driving decision. These complex visual environments introduce more irrelevant or redundant cues, which can easily mislead the model's attention and reasoning, ultimately leading to degraded performance.

Different decision tasks exhibit varying reliance on visual inputs versus textual inputs. For example, turning decisions depend primarily on navigation commands, where additional visual context may be redundant or even misleading. Conversely, for the speed control task, the model may benefit more from image inputs, which can provide insight into upcoming traffic patterns, enabling more informed decisions about acceleration or deceleration. This pattern is empirically supported by the results in Table 8, where 82.8%

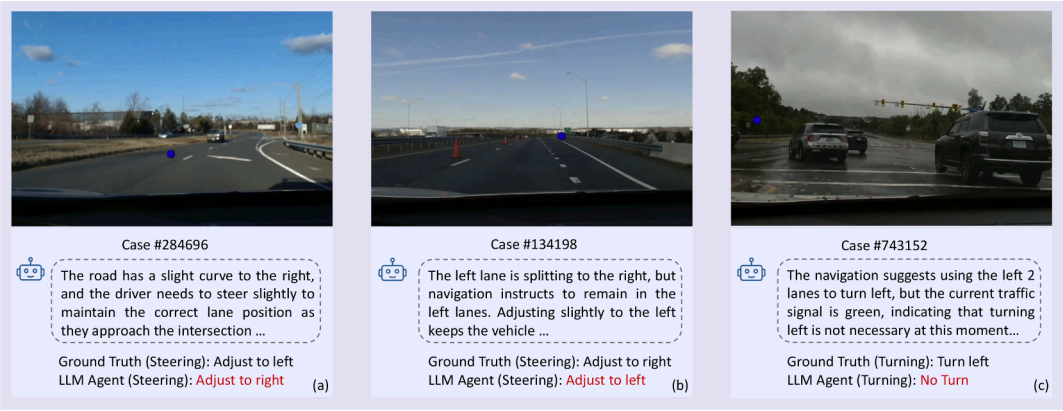


Fig. 11. Error cases where the inclusion of image inputs leads to degraded model performance in turning and steering tasks. (a) A spatial reasoning failure caused by complex road geometry. (b) A failure due to a mismatch between road topology and the navigation instruction, leading to incorrect steering. (c) A misinterprets of traffic signals resulting in a decision misaligned with the intended turn.

(9 out of 11) models exhibit improved accuracy when image input is included in the speed control task. In contrast, only 27.3% (3 of 11) models show improvement in the turning task, and 36.4% (4 of 11) in the steering task.

Fig. 11 illustrates representative cases of performance degradation observed in the turning and steering tasks, where models initially produced correct predictions under text-only conditions but failed after incorporating image inputs. These cases exemplify two common failure modes.

- **Incorrect judgment of surrounding vehicles and traffic signals** can arise when visual inputs are misinterpreted. Fig. 11 c presents an example at an intersection where the correct action, as indicated by the navigation instruction, is to make a left turn. However, the LLM incorrectly interprets the presence of a green traffic light as a signal to continue straight, ignoring the turning intent and producing a decision misaligned with that of the human driver.
- **Cross-modal conflicts in spatial reasoning** often occur when road geometry in the image contradicts navigation instructions. Fig. 11 a illustrates such a case involving a complex road geometry where a merge and a split occur within a short distance. The LLM, influenced by the visual cue of the road bending slightly to the right, incorrectly infers that a right steering action is appropriate. This prediction contradicts the human driver's left steering decision, which correctly aligns with the navigation intent and overall traffic context.

To provide more insights about this problem, we conduct a comprehensive error analysis across all evaluated models, focusing specifically on cases where the inclusion of image inputs leads to performance degradation. Building upon the three representative cases discussed in the original manuscript, we further refine image-induced decision errors into the following five categories:

- **Road geometry misread**: Errors caused by incorrect perception or understanding of road topology, such as misinterpreting lane curvature, ramp structures, merges, or road boundaries.
- **Signal agent misread**: Failures in correctly recognizing or reasoning about traffic signals, signs, or the intentions and states of surrounding agents (e.g., vehicles or pedestrians).
- **Visibility environment misread**: Errors arising from degraded environmental visibility or adverse conditions (e.g., low illumination, rain, fog), leading to incomplete or distorted visual perception.
- **Irrelevant saliency**: Cases where the model over-attends to visually salient but task-irrelevant elements, distracting the agent from critical navigation cues.
- **Navigation visual conflict**: Conflicts between visual inputs and navigation instructions or map-based priors, resulting in inconsistent or incorrect decision-making.
- **Other**: Rare or ambiguous failure cases that do not fall cleanly into the above categories.

The relationship between different models, error categories, and scenario types are illustrated by Fig. 12. Among all image-induced errors, the two most prevalent categories are Road Geometry Misread and Signal/Agent Misread, accounting for 19.1% and 18.8% of the total errors, respectively.

Notably, the dominant error types vary substantially across scenarios. In dusk conditions, errors caused by Road Geometry Misread (28.6%) and Visibility/Environment Misread (25.0%) become significantly more prominent, indicating that insufficient lighting leads to incorrect road structure perception and reduced visual clarity. In heavy rain and heavy traffic scenarios, Visibility/Environment Misread and Signal/Agent Misread dominate, suggesting that overly complex and noisy visual inputs impair the LLM agent's ability to make accurate judgments. Furthermore, in ramp scenarios, the proportion of errors attributed to Road Geometry Misread increases to 31.9%, highlighting the limitations of current vision-language models in understanding complex road geometries.

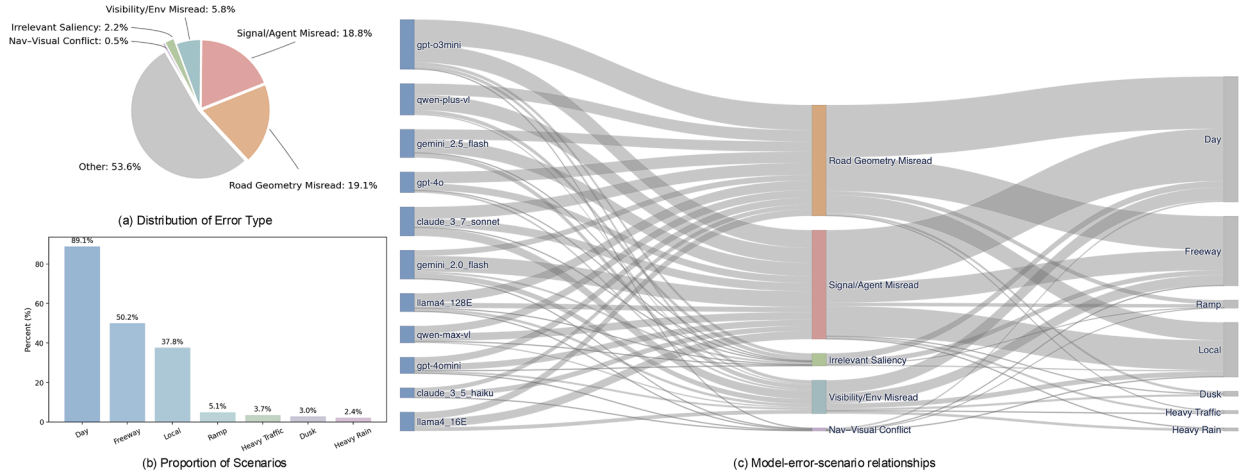


Fig. 12. Distribution of image-induced decision degradation across different multimodal LLM agents and driving scenarios. (a) Proportion of different image-induced error categories aggregated over all evaluated models.(b) Proportion of decision degradation occurrences under different driving scenario types. (c) Breakdown of error category distributions across scenarios and models, illustrating how dominant error types vary with environmental conditions and road structures.

5.6. *Visual ablation analysis.*

To further understand why the inclusion of image inputs improves performance in some cases while degrading it in others, we conduct controlled visual ablation studies that explicitly manipulate the spatial content of visual inputs, using two complementary strategies: gaze-based masking and masking of irrelevant image regions.

Gaze-based visual ablation. We formalize the gaze-based visual ablation as follows. Let I denote the original input image and G_t the gaze region accumulated over the t seconds preceding each key frame. We construct a binary gaze mask $M_t \in \{0,1\}^{H\times W}$ such that pixels within G_t are preserved and pixels outside G_t are suppressed. The ablated image is defined as

$$\tilde{I}_t = I \odot M_t,$$

where pixels outside the gaze region are set to a constant black value. We evaluate four temporal aggregation windows, $t \in \{1, 10, 20, 30\}$, to systematically examine the effect of gaze temporal extent on visual conditioning.

Masking irrelevant image regions. We formalize the irrelevant-region masking procedure as follows. Let $\mathcal{O}_{\text{scene}}$ denote the set of all objects present in the image, $\mathcal{O}_{\text{reason}}$ the set of visual objects referenced in the model's reasoning, and $\mathcal{O}_{\text{prompt}}$ the set of objects described in the textual prompt (e.g., environment and vehicle dynamics). The set of irrelevant objects is defined as

$$\mathcal{O}_i = \mathcal{O}_{\text{scene}} \setminus \left(\mathcal{O}_{\text{reason}} \cup \mathcal{O}_{\text{prompt}}\right).$$

To localize objects in $\mathcal{O}_i$, we employ an object detection model and retain detections with confidence scores above a threshold $\tau = 0.7$. For each detected object $o \in \mathcal{O}_i$, the corresponding bounding box region is replaced with a constant black mask, producing an ablated image $\tilde{I}$. This controlled masking removes visually present but semantically irrelevant content while preserving task-relevant scene information, enabling a quantitative evaluation of the impact of visual relevance on multimodal fusion and downstream turning and steering performance.

The quantitative results of the visual ablation experiments are summarized in Table 11. Overall, the impact of controlled visual ablation is strongly model-dependent and task-dependent, revealing important differences in how multimodal large language models exploit visual information for decision making.

- **Effect of gaze-based visual ablation.** For GPT-4o, gaze-based masking consistently degrades performance across most temporal windows, with overall accuracy dropping by 0.08 relative to the baseline. This indicates that GPT-4o relies on global scene context, and restricting visual input to localized gaze regions removes auxiliary cues that are implicitly leveraged during control reasoning. In contrast, Claude 3.7 sonnet benefits substantially from gaze-based visual conditioning. Gaze masking yields an overall accuracy improvement of up to 0.13, with the largest gains observed in Turn and Steer tasks. In particular, steering performance improves by up to 0.21 under longer temporal gaze aggregation. A similar trend is observed for Gemini 2.5 flash, where gaze-based masking improves overall accuracy by up to 0.14. Turn and Steer tasks again exhibit the strongest sensitivity, with improvements of up to 0.05 and 0.15, respectively. Across both Claude and Gemini, improvements are consistently larger for Turn and Steer than for Speed, confirming that steering-related tasks are more sensitive to localized spatial cues and structured visual input, whereas speed prediction relies more heavily on global scene dynamics.

Table 11
Visual ablation results under gaze-based masking and irrelevant-region masking. Results are reported for GPT-4o, Claude 3.7 Sonnet, and Gemini 2.5 Flash; Gaze Mask-t preserves only gaze-attended regions within the preceding t seconds. Accuracy is shown for Speed, Turn, and Steer tasks, with Overall denoting joint task performance.

Model	Visual Input	Task Accuracy			
		Speed	Turn	Steer	Overall
GPT-4o	Baseline	0.71	0.84	0.76	0.45
	Gaze Mask-1s	0.71	0.82	0.68	0.37
	Gaze Mask-10s	0.66	0.79	0.66	0.34
	Gaze Mask-20s	0.63	0.82	0.63	0.34
	Gaze Mask-30s	0.71	0.82	0.71	0.39
	Mask Irrelevant	0.58	0.76	0.70	0.26
Claude 3.7 Sonnet	Baseline	0.61	0.82	0.66	0.34
	Gaze Mask-1s	0.58	0.82	0.84	0.42
	Gaze Mask-10s	0.55	0.82	0.82	0.39
	Gaze Mask-20s	0.58	0.87	0.87	0.47
	Gaze Mask-30s	0.53	0.79	0.84	0.45
	Mask Irrelevant	0.50	0.82	0.76	0.22
Gemini 2.5 Flash	Baseline	0.71	0.84	0.61	0.39
	Gaze Mask-1s	0.74	0.87	0.74	0.47
	Gaze Mask-10s	0.61	0.87	0.79	0.45
	Gaze Mask-20s	0.68	0.89	0.76	0.53
	Gaze Mask-30s	0.63	0.89	0.68	0.37
	Mask Irrelevant	0.56	0.82	0.72	0.34

- **Effect of masking irrelevant image regions.** In contrast to gaze-based masking, masking irrelevant image regions results in consistent performance degradation across all models. Overall accuracy decreases 0.19 for GPT-4o, 0.12 for Claude 3.7 sonnet, and 0.05 for Gemini 2.5 flash, indicating that object-level removal is broadly detrimental to multimodal decision making. This degradation suggests that machine-defined semantic irrelevance may not align with decision relevance in driving scenarios. Objects excluded based on explicit reasoning or prompt semantics may still contribute essential structural context. Removing these regions disrupts the global visual distribution, impairing multimodal fusion and leading to degraded on multiple control performance.

These results highlight a key distinction between attention-aligned visual filtering and machine perceived object-level filtering. Gaze-based ablation, which preserves human-attended regions, can significantly improve steering and turning performance. In contrast, masking semantically "irrelevant" objects consistently harms performance, revealing a mismatch between machine-defined visual relevance and the latent visual cues exploited during decision making.

6. Future works and limitations

Building on our current dataset and evaluation methodology, several promising directions emerge for future exploration. First, the dataset can be used to fine-tune LLM agents that embody a broad range of human driving profiles, thereby introducing greater behavioral diversity and realism into safety-focused traffic simulations. Beyond the HermesSense, future work can also investigate how various input streams, such as visual cues, sensor data, and contextual information collectively inform driving decisions. These insights pave the way for designing novel agent reasoning processes that are both more interpretable and more closely aligned with human behavior. Furthermore, deeper investigations into how information flows through decision pipelines may inspire innovative model architectures capable of robust decision-making under dynamic traffic conditions. By mapping model outputs back to specific input signals, our framework sheds light on the reasoning processes that guide LLM decision-making in complex, real-world scenarios. This added interpretability not only helps diagnose incorrect or unsafe driving behaviors but also promotes accountability for traffic safety.

Limitations. Despite advancements, the framework may still struggle with sudden hazards or rapidly shifting behaviors. Additionally, its reliance on U.S. data risks underrepresenting certain driver demographics and driving scenarios from other countries.

7. Conclustion

In this paper, we present HermesSense, a human-centric multimodal dataset that integrates both sensory and semantic information for autonomous driving. Built upon this dataset, we introduce a comprehensive evaluation framework to assess the human-likeness of LLM-based driving agents, pecifically, from behavioral alignment and statistical human-likeness, capturing both short-term decision consistency and long-term behavioral patterns. This evaluation offers a rigorous foundation for quantifying how well LLM agents replicate human driving behavior, laying the groundwork for the development of future autonomous systems that are capable of

adapting to the dynamic, interactive nature of real-world traffic. Beyond the HermesSense, our data attribution framework, which pinpoints the most influential input components shaping human-like behavior, holds clear potential for advancing autonomous driving and robotic tasks.

CRediT authorship contribution statement

Pu Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization; **Guanlin Wu:** Writing – review & editing, Visualization, Data curation; **Yang Zhao:** Writing – review & editing, Methodology, Formal analysis, Conceptualization; **Yichen Lin:** Writing – review & editing, Visualization, Data curation; **Hao Frank Yang:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

None.

Appendix A. Appendix

A.1. Prompt examples

Fig. A.1 illustrate two prompt examples from the HermesSense dataset. These prompts serve as structured textual inputs to the LLM-based agent, encoding multimodal driving context to facilitate reasoning and behavior prediction.

A.2. Ablation study on effect of sequence length

We report the detailed $\mathbb{B}_{\text{sim}}$ values and their variances across the five driver demographic groups shown in Tables A.1–A.5.

Example of organized prompt for #208

Driving Scenario Description

[Driver demographic information]:
Assume you are a 27-year-old Asian male driver with 3 years of driving experience, driving approximately 150 miles per week. You will be provided with descriptions of real-world driving scenarios. Based on these details, simulate the driver's decision-making and behavior.

[Environment information]:
The driving environment is on a straight highway with multiple lanes, under overcast skies transitioning to dusk. The road is flat with no visible inclines, curves, or elevation changes. Lane markings are clear and include solid and dashed lines. The road surface appears smooth without any noticeable damage or debris. Overhead signage indicates upcoming exits and directions. There are no visible intersections, pedestrian crossings, or construction zones. Traffic signals are absent on this stretch of highway. Barriers separate the lanes from oncoming traffic. Traffic density is moderate with a steady flow, and there is consistent movement across the lanes. roadside trees and power lines are visible.

[Vehicle dynamics & Driver behavior information]:
The vehicle is maintaining a steady speed of 111 kilometers per hour. The driver's throttle input over 3 seconds was ['low', 'low', 'low'], and the brake input was ['none', 'none', 'none'].

[Traffic dynamics information]:
There are 4 lanes on the road. The ego vehicle is driving in the second lane from the left.
In the leftmost lane, there is no vehicle. In the second lane from the left, there is a vehicle at Far distance, gap is maintaining distance. In the second lane from the right, there is a vehicle at Close distance, the gap is closing. In the rightmost lane, there is no vehicle.

[Navigation information]:
Navigation indicates the optimal route is to stay on the main road.

[Gaze information]:
In the highlighted regions of the driver-view images, a multi-lane road is visible, with a clear sky filled with scattered clouds. In the foreground, a barrier separates the road from the surrounding landscape, which includes trees. Several red vehicles are seen traveling in the same direction as the driver. A traffic light is visible in the distance, indicating an intersection ahead.

Task Instruction

Based on the provided driving scenario, select the most likely driver action at each step. After selecting, provide a reason for your choice based on road conditions, traffic flow, and driving safety considerations.
Step 1: Speed Adjustment
- Maintain speed - Accelerate - Decelerate
Step 2: Turning Decision (This decision is made specifically at intersections and does not occur simultaneously with a steering change.)
- Turn left - Turn right - No turn
Step 3: Steering Change Decision (Including lane change maneuvers and evasive steering.)
- Adjust the steering to left - Adjust the steering to right - Maintain the steering

---

The ground truth is: *<Maintain speed>, <No turn>, <Maintain the steering>*

Fig. A.1. A prompt example for #208 instance.

Table A.1
Variance of B_{sim} across different sequence length L for the *has accident* category. Results are reported as mean ± standard deviation. The subscript I denotes models that use images as input.

Model	$L = 10$	$L = 20$	$L = 30$	$L = 40$	$L = 50$	$L = 60$
Has accident - No						
claude-3-5-haiku	0.780 ± 0.040	0.778 ± 0.024	0.774 ± 0.034	0.776 ± 0.026	0.770 ± 0.022	0.769 ± 0.024
claude-3-5-haiku$_I$	0.784 ± 0.045	0.781 ± 0.033	0.771 ± 0.038	0.779 ± 0.029	0.780 ± 0.022	0.774 ± 0.021
claude-3-7-sonnet	0.816 ± 0.044	0.794 ± 0.041	0.799 ± 0.035	0.810 ± 0.026	0.799 ± 0.025	0.810 ± 0.024
claude-3-7-sonnet$_I$	0.813 ± 0.045	0.810 ± 0.036	0.804 ± 0.025	0.799 ± 0.030	0.806 ± 0.030	0.797 ± 0.031
gemini-2.0-flash	0.802 ± 0.039	0.797 ± 0.036	0.795 ± 0.039	0.796 ± 0.021	0.792 ± 0.031	0.801 ± 0.025
gemini-2.0-flash$_I$	0.780 ± 0.055	0.787 ± 0.050	0.786 ± 0.037	0.775 ± 0.032	0.790 ± 0.031	0.791 ± 0.025
gemini-2.5-flash	0.776 ± 0.041	0.789 ± 0.032	0.778 ± 0.030	0.764 ± 0.031	0.780 ± 0.028	0.772 ± 0.032
gemini-2.5-flash$_I$	0.773 ± 0.051	0.780 ± 0.047	0.787 ± 0.034	0.783 ± 0.027	0.777 ± 0.032	0.781 ± 0.025
gpt-4o	0.800 ± 0.066	0.792 ± 0.037	0.794 ± 0.050	0.780 ± 0.031	0.785 ± 0.029	0.796 ± 0.023
gpt-4o$_I$	0.794 ± 0.057	0.800 ± 0.052	0.806 ± 0.027	0.793 ± 0.036	0.804 ± 0.033	0.803 ± 0.032
gpt-4omini	0.758 ± 0.054	0.763 ± 0.042	0.766 ± 0.036	0.767 ± 0.040	0.762 ± 0.026	0.761 ± 0.030
gpt-4omini$_I$	0.772 ± 0.055	0.775 ± 0.041	0.777 ± 0.037	0.770 ± 0.033	0.778 ± 0.029	0.755 ± 0.022
gpt-o3mini	0.796 ± 0.051	0.786 ± 0.044	0.788 ± 0.032	0.786 ± 0.032	0.795 ± 0.028	0.785 ± 0.027
gpt-o3mini$_I$	0.762 ± 0.043	0.770 ± 0.038	0.759 ± 0.040	0.763 ± 0.028	0.756 ± 0.035	0.768 ± 0.027
llama4-128E	0.783 ± 0.032	0.780 ± 0.034	0.774 ± 0.033	0.772 ± 0.030	0.775 ± 0.023	0.767 ± 0.029
llama4-128E$_I$	0.785 ± 0.052	0.773 ± 0.034	0.791 ± 0.030	0.780 ± 0.025	0.777 ± 0.025	0.777 ± 0.028
llama4-16E	0.797 ± 0.054	0.792 ± 0.036	0.786 ± 0.036	0.783 ± 0.030	0.794 ± 0.032	0.789 ± 0.026
llama4-16E$_I$	0.820 ± 0.046	0.815 ± 0.038	0.793 ± 0.039	0.794 ± 0.031	0.797 ± 0.031	0.804 ± 0.025
qwen-max-vl	0.791 ± 0.051	0.769 ± 0.042	0.789 ± 0.043	0.782 ± 0.032	0.784 ± 0.029	0.782 ± 0.029
qwen-max-vl$_I$	0.802 ± 0.053	0.783 ± 0.046	0.790 ± 0.038	0.791 ± 0.040	0.782 ± 0.036	0.793 ± 0.040
qwen-plus-vl	0.803 ± 0.037	0.791 ± 0.038	0.786 ± 0.035	0.790 ± 0.027	0.793 ± 0.025	0.793 ± 0.024
qwen-plus-vl$_I$	0.772 ± 0.045	0.771 ± 0.031	0.761 ± 0.034	0.766 ± 0.034	0.774 ± 0.024	0.773 ± 0.029
Has accident - Yes						
claude-3-5-haiku	0.785 ± 0.052	0.797 ± 0.024	0.802 ± 0.029	0.801 ± 0.016	0.804 ± 0.016	0.802 ± 0.015
claude-3-5-haiku$_I$	0.805 ± 0.046	0.812 ± 0.024	0.804 ± 0.018	0.800 ± 0.019	0.808 ± 0.017	0.809 ± 0.009
claude-3-7-sonnet	0.827 ± 0.056	0.834 ± 0.033	0.834 ± 0.042	0.827 ± 0.027	0.826 ± 0.027	0.832 ± 0.026
claude-3-7-sonnet$_I$	0.852 ± 0.053	0.848 ± 0.038	0.850 ± 0.033	0.842 ± 0.030	0.839 ± 0.026	0.850 ± 0.016
gemini-2.0-flash	0.806 ± 0.064	0.843 ± 0.023	0.826 ± 0.034	0.833 ± 0.026	0.837 ± 0.019	0.833 ± 0.018
gemini-2.0-flash$_I$	0.808 ± 0.041	0.817 ± 0.034	0.814 ± 0.028	0.810 ± 0.026	0.814 ± 0.020	0.811 ± 0.015
gemini-2.5-flash	0.810 ± 0.044	0.815 ± 0.034	0.811 ± 0.029	0.816 ± 0.029	0.816 ± 0.025	0.813 ± 0.017
gemini-2.5-flash$_I$	0.809 ± 0.053	0.813 ± 0.034	0.826 ± 0.027	0.820 ± 0.030	0.829 ± 0.023	0.824 ± 0.016
gpt-4o	0.816 ± 0.042	0.821 ± 0.030	0.822 ± 0.030	0.819 ± 0.024	0.815 ± 0.014	0.816 ± 0.016
gpt-4o$_I$	0.811 ± 0.046	0.814 ± 0.035	0.823 ± 0.028	0.824 ± 0.022	0.828 ± 0.017	0.823 ± 0.015
gpt-4omini	0.804 ± 0.045	0.797 ± 0.035	0.792 ± 0.037	0.798 ± 0.030	0.803 ± 0.020	0.797 ± 0.021
gpt-4omini$_I$	0.801 ± 0.046	0.798 ± 0.033	0.798 ± 0.025	0.792 ± 0.017	0.800 ± 0.015	0.802 ± 0.012
gpt-o3mini	0.841 ± 0.053	0.823 ± 0.028	0.827 ± 0.034	0.818 ± 0.020	0.823 ± 0.018	0.822 ± 0.019
gpt-o3mini$_I$	0.791 ± 0.050	0.801 ± 0.034	0.792 ± 0.026	0.797 ± 0.023	0.798 ± 0.018	0.797 ± 0.012
llama4-128E	0.799 ± 0.064	0.811 ± 0.044	0.808 ± 0.046	0.810 ± 0.044	0.811 ± 0.033	0.817 ± 0.025
llama4-128E$_I$	0.819 ± 0.061	0.832 ± 0.049	0.822 ± 0.042	0.821 ± 0.035	0.827 ± 0.040	0.817 ± 0.032
llama4-16E	0.826 ± 0.043	0.820 ± 0.035	0.839 ± 0.024	0.832 ± 0.030	0.831 ± 0.019	0.830 ± 0.022
llama4-16E$_I$	0.834 ± 0.047	0.836 ± 0.034	0.838 ± 0.020	0.841 ± 0.018	0.839 ± 0.020	0.836 ± 0.021
qwen-max-vl	0.832 ± 0.049	0.838 ± 0.034	0.820 ± 0.035	0.833 ± 0.025	0.839 ± 0.023	0.834 ± 0.023
qwen-max-vl$_I$	0.823 ± 0.057	0.839 ± 0.035	0.833 ± 0.036	0.826 ± 0.027	0.826 ± 0.020	0.836 ± 0.021
qwen-plus-vl	0.801 ± 0.036	0.816 ± 0.030	0.807 ± 0.023	0.815 ± 0.017	0.812 ± 0.019	0.819 ± 0.014
qwen-plus-vl$_I$	0.780 ± 0.037	0.779 ± 0.038	0.780 ± 0.028	0.783 ± 0.024	0.785 ± 0.021	0.780 ± 0.017

Table A.2
Variance of B_{sim} across different sequence length L for the *driver age* category. Results are reported as mean ± standard deviation. The subscript I denotes models that use images as input.

Model	$L = 10$	$L = 20$	$L = 30$	$L = 40$	$L = 50$	$L = 60$
Driver age - <=25 old						
claude-3-5-haiku	0.777 ± 0.056	0.802 ± 0.021	0.800 ± 0.026	0.800 ± 0.014	0.796 ± 0.015	0.798 ± 0.012
claude-3-5-haiku$_I$	0.784 ± 0.045	0.797 ± 0.021	0.788 ± 0.025	0.799 ± 0.019	0.795 ± 0.017	0.792 ± 0.015
claude-3-7-sonnet	0.838 ± 0.039	0.821 ± 0.033	0.826 ± 0.022	0.817 ± 0.026	0.813 ± 0.023	0.821 ± 0.020
claude-3-7-sonnet$_I$	0.826 ± 0.048	0.847 ± 0.039	0.830 ± 0.033	0.832 ± 0.029	0.835 ± 0.018	0.835 ± 0.020
gemini-2.0-flash	0.834 ± 0.043	0.830 ± 0.030	0.837 ± 0.028	0.837 ± 0.017	0.837 ± 0.018	0.835 ± 0.015
gemini-2.0-flash$_I$	0.813 ± 0.046	0.825 ± 0.043	0.822 ± 0.035	0.827 ± 0.022	0.831 ± 0.017	0.818 ± 0.020
gemini-2.5-flash	0.803 ± 0.046	0.791 ± 0.044	0.800 ± 0.037	0.801 ± 0.029	0.803 ± 0.025	0.800 ± 0.018
gemini-2.5-flash$_I$	0.818 ± 0.054	0.826 ± 0.024	0.818 ± 0.020	0.819 ± 0.019	0.818 ± 0.018	0.815 ± 0.015
gpt-4o	0.828 ± 0.054	0.810 ± 0.042	0.828 ± 0.024	0.824 ± 0.025	0.823 ± 0.022	0.820 ± 0.019
gpt-4o$_I$	0.835 ± 0.051	0.825 ± 0.036	0.836 ± 0.031	0.834 ± 0.021	0.825 ± 0.020	0.829 ± 0.015
gpt-4omini	0.802 ± 0.035	0.783 ± 0.048	0.816 ± 0.024	0.803 ± 0.027	0.803 ± 0.021	0.802 ± 0.021
gpt-4omini$_I$	0.802 ± 0.045	0.798 ± 0.042	0.806 ± 0.024	0.807 ± 0.022	0.803 ± 0.018	0.804 ± 0.013
gpt-o3mini	0.818 ± 0.039	0.813 ± 0.050	0.825 ± 0.026	0.826 ± 0.019	0.820 ± 0.018	0.825 ± 0.015
gpt-o3mini$_I$	0.792 ± 0.050	0.804 ± 0.041	0.808 ± 0.023	0.799 ± 0.022	0.808 ± 0.014	0.805 ± 0.015
llama4-128E	0.819 ± 0.045	0.798 ± 0.039	0.806 ± 0.030	0.806 ± 0.023	0.802 ± 0.017	0.805 ± 0.020
llama4-128E$_I$	0.802 ± 0.052	0.799 ± 0.035	0.805 ± 0.039	0.806 ± 0.037	0.809 ± 0.022	0.813 ± 0.025
llama4-16E	0.838 ± 0.037	0.832 ± 0.035	0.830 ± 0.034	0.838 ± 0.020	0.835 ± 0.013	0.833 ± 0.017
llama4-16E$_I$	0.841 ± 0.028	0.836 ± 0.026	0.842 ± 0.020	0.838 ± 0.016	0.839 ± 0.014	0.837 ± 0.010
qwen-max-vl	0.805 ± 0.044	0.814 ± 0.039	0.815 ± 0.033	0.824 ± 0.030	0.815 ± 0.025	0.816 ± 0.023
qwen-max-vl$_I$	0.823 ± 0.038	0.825 ± 0.040	0.820 ± 0.030	0.821 ± 0.024	0.821 ± 0.017	0.819 ± 0.011
qwen-plus-vl	0.811 ± 0.029	0.817 ± 0.024	0.822 ± 0.015	0.816 ± 0.013	0.818 ± 0.010	0.815 ± 0.009
qwen-plus-vl$_I$	0.792 ± 0.036	0.796 ± 0.032	0.798 ± 0.020	0.793 ± 0.019	0.796 ± 0.013	0.794 ± 0.012
Driver age - >25 old						
claude-3-5-haiku	0.772 ± 0.040	0.768 ± 0.045	0.773 ± 0.030	0.774 ± 0.028	0.756 ± 0.030	0.770 ± 0.028
claude-3-5-haiku$_I$	0.767 ± 0.045	0.783 ± 0.039	0.779 ± 0.030	0.775 ± 0.031	0.778 ± 0.022	0.774 ± 0.024
claude-3-7-sonnet	0.799 ± 0.057	0.801 ± 0.043	0.816 ± 0.041	0.801 ± 0.031	0.807 ± 0.028	0.799 ± 0.032
claude-3-7-sonnet$_I$	0.803 ± 0.068	0.804 ± 0.042	0.800 ± 0.041	0.803 ± 0.029	0.806 ± 0.035	0.810 ± 0.034
gemini-2.0-flash	0.782 ± 0.068	0.790 ± 0.046	0.784 ± 0.035	0.790 ± 0.029	0.794 ± 0.031	0.795 ± 0.030
gemini-2.0-flash$_I$	0.795 ± 0.050	0.790 ± 0.038	0.772 ± 0.041	0.774 ± 0.028	0.787 ± 0.026	0.780 ± 0.030
gemini-2.5-flash	0.792 ± 0.067	0.780 ± 0.038	0.799 ± 0.034	0.786 ± 0.032	0.786 ± 0.035	0.790 ± 0.032
gemini-2.5-flash$_I$	0.776 ± 0.053	0.781 ± 0.057	0.792 ± 0.039	0.789 ± 0.032	0.787 ± 0.030	0.800 ± 0.027
gpt-4o	0.796 ± 0.033	0.790 ± 0.039	0.792 ± 0.035	0.790 ± 0.033	0.793 ± 0.031	0.793 ± 0.022
gpt-4o$_I$	0.802 ± 0.051	0.794 ± 0.042	0.811 ± 0.035	0.794 ± 0.038	0.800 ± 0.030	0.803 ± 0.025
gpt-4omini	0.780 ± 0.051	0.785 ± 0.030	0.761 ± 0.032	0.770 ± 0.032	0.763 ± 0.031	0.768 ± 0.022
gpt-4omini$_I$	0.749 ± 0.048	0.771 ± 0.034	0.767 ± 0.027	0.782 ± 0.033	0.767 ± 0.029	0.772 ± 0.025
gpt-o3mini	0.791 ± 0.053	0.819 ± 0.035	0.811 ± 0.024	0.798 ± 0.032	0.793 ± 0.026	0.800 ± 0.031
gpt-o3mini$_I$	0.761 ± 0.054	0.759 ± 0.037	0.761 ± 0.032	0.774 ± 0.025	0.767 ± 0.026	0.756 ± 0.022
llama4-128E	0.759 ± 0.051	0.783 ± 0.053	0.772 ± 0.043	0.773 ± 0.038	0.774 ± 0.038	0.778 ± 0.030
llama4-128E$_I$	0.778 ± 0.061	0.789 ± 0.040	0.790 ± 0.039	0.785 ± 0.033	0.793 ± 0.024	0.788 ± 0.020
llama4-16E	0.770 ± 0.065	0.787 ± 0.038	0.799 ± 0.034	0.800 ± 0.040	0.786 ± 0.033	0.792 ± 0.029
llama4-16E$_I$	0.806 ± 0.049	0.794 ± 0.040	0.802 ± 0.034	0.814 ± 0.036	0.798 ± 0.040	0.793 ± 0.026
qwen-max-vl	0.785 ± 0.052	0.797 ± 0.039	0.789 ± 0.043	0.796 ± 0.046	0.804 ± 0.033	0.796 ± 0.038
qwen-max-vl$_I$	0.774 ± 0.055	0.784 ± 0.043	0.786 ± 0.042	0.796 ± 0.045	0.781 ± 0.036	0.795 ± 0.035
qwen-plus-vl	0.802 ± 0.045	0.788 ± 0.034	0.789 ± 0.035	0.793 ± 0.033	0.788 ± 0.024	0.801 ± 0.022
qwen-plus-vl$_I$	0.754 ± 0.039	0.762 ± 0.034	0.773 ± 0.034	0.767 ± 0.025	0.770 ± 0.028	0.766 ± 0.027

Table A.3
Variance of B_{sim} across different sequence length L for the *driving style* category. Results are reported as mean ± standard deviation. The subscript I denotes models that use images as input.

Model	$L = 10$	$L = 20$	$L = 30$	$L = 40$	$L = 50$	$L = 60$
Driver gender - Female						
claude-3-5-haiku	0.788 ± 0.041	0.796 ± 0.041	0.782 ± 0.032	0.790 ± 0.030	0.793 ± 0.024	0.791 ± 0.025
claude-3-5-haiku$_I$	0.792 ± 0.042	0.790 ± 0.032	0.788 ± 0.036	0.792 ± 0.026	0.783 ± 0.028	0.792 ± 0.028
claude-3-7-sonnet	0.822 ± 0.035	0.814 ± 0.029	0.814 ± 0.028	0.824 ± 0.030	0.817 ± 0.024	0.822 ± 0.017
claude-3-7-sonnet$_I$	0.821 ± 0.042	0.813 ± 0.048	0.822 ± 0.040	0.823 ± 0.031	0.827 ± 0.029	0.823 ± 0.025
gemini-2.0-flash	0.808 ± 0.045	0.803 ± 0.039	0.796 ± 0.031	0.804 ± 0.031	0.803 ± 0.027	0.804 ± 0.031
gemini-2.0-flash$_I$	0.790 ± 0.042	0.780 ± 0.052	0.793 ± 0.030	0.798 ± 0.027	0.797 ± 0.034	0.799 ± 0.035
gemini-2.5-flash	0.805 ± 0.057	0.794 ± 0.044	0.795 ± 0.035	0.798 ± 0.035	0.795 ± 0.031	0.791 ± 0.025
gemini-2.5-flash$_I$	0.795 ± 0.054	0.811 ± 0.043	0.804 ± 0.035	0.789 ± 0.028	0.802 ± 0.029	0.802 ± 0.020
gpt-4o	0.813 ± 0.056	0.808 ± 0.036	0.808 ± 0.032	0.817 ± 0.028	0.811 ± 0.028	0.804 ± 0.029
gpt-4o$_I$	0.807 ± 0.065	0.817 ± 0.039	0.809 ± 0.044	0.826 ± 0.028	0.821 ± 0.030	0.809 ± 0.033
gpt-4omini	0.774 ± 0.055	0.783 ± 0.040	0.782 ± 0.033	0.783 ± 0.030	0.791 ± 0.029	0.785 ± 0.031
gpt-4omini$_I$	0.780 ± 0.047	0.773 ± 0.035	0.783 ± 0.027	0.785 ± 0.032	0.778 ± 0.025	0.787 ± 0.027
gpt-o3mini	0.812 ± 0.048	0.813 ± 0.026	0.797 ± 0.031	0.815 ± 0.026	0.806 ± 0.026	0.805 ± 0.023
gpt-o3mini$_I$	0.772 ± 0.055	0.789 ± 0.039	0.778 ± 0.042	0.786 ± 0.026	0.783 ± 0.031	0.782 ± 0.034
llama4-128E	0.801 ± 0.045	0.785 ± 0.041	0.793 ± 0.038	0.793 ± 0.029	0.798 ± 0.028	0.786 ± 0.026
llama4-128E$_I$	0.801 ± 0.045	0.809 ± 0.042	0.806 ± 0.042	0.795 ± 0.027	0.798 ± 0.030	0.800 ± 0.027
llama4-16E	0.804 ± 0.058	0.810 ± 0.045	0.815 ± 0.037	0.815 ± 0.040	0.807 ± 0.035	0.801 ± 0.036
llama4-16E$_I$	0.814 ± 0.058	0.802 ± 0.041	0.820 ± 0.035	0.829 ± 0.028	0.814 ± 0.032	0.822 ± 0.028
qwen-max-vl	0.796 ± 0.062	0.794 ± 0.042	0.799 ± 0.047	0.798 ± 0.041	0.797 ± 0.046	0.792 ± 0.035
qwen-max-vl$_I$	0.791 ± 0.052	0.799 ± 0.050	0.805 ± 0.048	0.789 ± 0.042	0.809 ± 0.043	0.792 ± 0.043
qwen-plus-vl	0.792 ± 0.040	0.794 ± 0.041	0.793 ± 0.035	0.793 ± 0.030	0.811 ± 0.027	0.812 ± 0.024
qwen-plus-vl$_I$	0.760 ± 0.055	0.769 ± 0.030	0.753 ± 0.039	0.775 ± 0.028	0.759 ± 0.031	0.768 ± 0.033
Driver gender - Male						
claude-3-5-haiku	0.760 ± 0.048	0.775 ± 0.035	0.757 ± 0.036	0.772 ± 0.025	0.774 ± 0.026	0.768 ± 0.026
claude-3-5-haiku$_I$	0.777 ± 0.056	0.773 ± 0.047	0.770 ± 0.035	0.772 ± 0.029	0.769 ± 0.026	0.772 ± 0.025
claude-3-7-sonnet	0.782 ± 0.065	0.789 ± 0.048	0.806 ± 0.030	0.805 ± 0.038	0.801 ± 0.019	0.811 ± 0.028
claude-3-7-sonnet$_I$	0.805 ± 0.049	0.803 ± 0.046	0.809 ± 0.029	0.811 ± 0.035	0.806 ± 0.029	0.809 ± 0.030
gemini-2.0-flash	0.785 ± 0.051	0.796 ± 0.037	0.802 ± 0.025	0.803 ± 0.040	0.792 ± 0.032	0.804 ± 0.033
gemini-2.0-flash$_I$	0.798 ± 0.040	0.769 ± 0.047	0.790 ± 0.038	0.796 ± 0.038	0.792 ± 0.030	0.781 ± 0.034
gemini-2.5-flash	0.783 ± 0.064	0.795 ± 0.041	0.779 ± 0.042	0.778 ± 0.037	0.779 ± 0.032	0.784 ± 0.030
gemini-2.5-flash$_I$	0.788 ± 0.063	0.799 ± 0.048	0.786 ± 0.046	0.787 ± 0.043	0.789 ± 0.040	0.789 ± 0.031
gpt-4o	0.792 ± 0.054	0.791 ± 0.047	0.793 ± 0.035	0.803 ± 0.035	0.785 ± 0.035	0.781 ± 0.029
gpt-4o$_I$	0.797 ± 0.050	0.803 ± 0.030	0.786 ± 0.038	0.809 ± 0.030	0.794 ± 0.032	0.806 ± 0.034
gpt-4omini	0.770 ± 0.056	0.778 ± 0.050	0.766 ± 0.037	0.757 ± 0.039	0.766 ± 0.030	0.770 ± 0.032
gpt-4omini$_I$	0.764 ± 0.052	0.799 ± 0.037	0.776 ± 0.031	0.758 ± 0.031	0.764 ± 0.032	0.780 ± 0.027
gpt-o3mini	0.803 ± 0.046	0.796 ± 0.040	0.796 ± 0.038	0.805 ± 0.027	0.789 ± 0.034	0.805 ± 0.026
gpt-o3mini$_I$	0.776 ± 0.052	0.762 ± 0.044	0.762 ± 0.032	0.777 ± 0.033	0.772 ± 0.029	0.768 ± 0.021
llama4-128E	0.766 ± 0.051	0.778 ± 0.059	0.771 ± 0.041	0.772 ± 0.040	0.770 ± 0.038	0.777 ± 0.026
llama4-128E$_I$	0.785 ± 0.047	0.789 ± 0.035	0.790 ± 0.038	0.775 ± 0.031	0.785 ± 0.035	0.785 ± 0.028
llama4-16E	0.775 ± 0.060	0.792 ± 0.039	0.801 ± 0.035	0.794 ± 0.035	0.789 ± 0.032	0.793 ± 0.034
llama4-16E$_I$	0.791 ± 0.052	0.795 ± 0.039	0.805 ± 0.030	0.811 ± 0.026	0.800 ± 0.035	0.798 ± 0.033
qwen-max-vl	0.801 ± 0.053	0.798 ± 0.057	0.797 ± 0.041	0.794 ± 0.037	0.798 ± 0.038	0.806 ± 0.027
qwen-max-vl$_I$	0.779 ± 0.061	0.789 ± 0.048	0.796 ± 0.046	0.787 ± 0.045	0.806 ± 0.036	0.790 ± 0.037
qwen-plus-vl	0.802 ± 0.037	0.799 ± 0.034	0.789 ± 0.033	0.806 ± 0.022	0.797 ± 0.028	0.797 ± 0.030
qwen-plus-vl$_I$	0.783 ± 0.041	0.769 ± 0.041	0.768 ± 0.036	0.777 ± 0.026	0.777 ± 0.022	0.776 ± 0.028

Table A.4
Variance of B_{sim} across different sequence length L for the *driving years* category. Results are reported as mean ± standard deviation. The subscript I denotes models that use images as input.

Model	$L = 10$	$L = 20$	$L = 30$	$L = 40$	$L = 50$	$L = 60$
Driving style - Aggressive						
claude-3-5-haiku	0.798 ± 0.049	0.805 ± 0.031	0.807 ± 0.027	0.804 ± 0.015	0.803 ± 0.011	0.802 ± 0.014
claude-3-5-haiku$_I$	0.807 ± 0.052	0.803 ± 0.032	0.799 ± 0.032	0.802 ± 0.019	0.806 ± 0.016	0.809 ± 0.012
claude-3-7-sonnet	0.824 ± 0.063	0.820 ± 0.053	0.828 ± 0.035	0.831 ± 0.033	0.842 ± 0.027	0.828 ± 0.024
claude-3-7-sonnet$_I$	0.823 ± 0.064	0.834 ± 0.047	0.843 ± 0.033	0.850 ± 0.026	0.851 ± 0.026	0.841 ± 0.026
gemini-2.0-flash	0.842 ± 0.038	0.811 ± 0.040	0.831 ± 0.028	0.836 ± 0.027	0.831 ± 0.027	0.834 ± 0.022
gemini-2.0-flash$_I$	0.805 ± 0.054	0.808 ± 0.040	0.809 ± 0.034	0.817 ± 0.020	0.817 ± 0.018	0.815 ± 0.015
gemini-2.5-flash	0.812 ± 0.057	0.808 ± 0.053	0.815 ± 0.046	0.813 ± 0.039	0.816 ± 0.034	0.813 ± 0.026
gemini-2.5-flash$_I$	0.809 ± 0.059	0.829 ± 0.032	0.826 ± 0.032	0.821 ± 0.028	0.826 ± 0.025	0.824 ± 0.022
gpt-4o	0.802 ± 0.050	0.815 ± 0.039	0.813 ± 0.028	0.814 ± 0.023	0.811 ± 0.019	0.821 ± 0.016
gpt-4o$_I$	0.812 ± 0.061	0.834 ± 0.033	0.825 ± 0.030	0.824 ± 0.030	0.829 ± 0.018	0.828 ± 0.015
gpt-4omini	0.782 ± 0.059	0.788 ± 0.046	0.778 ± 0.040	0.792 ± 0.027	0.796 ± 0.024	0.799 ± 0.023
gpt-4omini$_I$	0.798 ± 0.044	0.792 ± 0.039	0.796 ± 0.023	0.799 ± 0.023	0.796 ± 0.018	0.802 ± 0.018
gpt-o3mini	0.808 ± 0.055	0.823 ± 0.035	0.820 ± 0.030	0.825 ± 0.028	0.814 ± 0.029	0.822 ± 0.015
gpt-o3mini$_I$	0.783 ± 0.056	0.802 ± 0.038	0.793 ± 0.028	0.801 ± 0.019	0.792 ± 0.020	0.798 ± 0.015
llama4-128E	0.809 ± 0.062	0.810 ± 0.056	0.812 ± 0.044	0.820 ± 0.035	0.812 ± 0.042	0.800 ± 0.032
llama4-128E$_I$	0.808 ± 0.064	0.817 ± 0.047	0.819 ± 0.045	0.822 ± 0.045	0.821 ± 0.042	0.830 ± 0.039
llama4-16E	0.836 ± 0.052	0.833 ± 0.032	0.821 ± 0.037	0.828 ± 0.028	0.824 ± 0.022	0.833 ± 0.022
llama4-16E$_I$	0.838 ± 0.040	0.834 ± 0.032	0.830 ± 0.035	0.833 ± 0.022	0.826 ± 0.022	0.835 ± 0.020
qwen-max-vl	0.828 ± 0.058	0.823 ± 0.040	0.831 ± 0.033	0.831 ± 0.026	0.836 ± 0.026	0.837 ± 0.026
qwen-max-vl$_I$	0.817 ± 0.058	0.826 ± 0.034	0.830 ± 0.032	0.837 ± 0.022	0.820 ± 0.024	0.832 ± 0.021
qwen-plus-vl	0.814 ± 0.029	0.820 ± 0.024	0.811 ± 0.020	0.813 ± 0.018	0.813 ± 0.018	0.822 ± 0.015
qwen-plus-vl$_I$	0.781 ± 0.042	0.785 ± 0.024	0.789 ± 0.023	0.789 ± 0.022	0.785 ± 0.017	0.790 ± 0.016
Driving style - Defensive						
claude-3-5-haiku	0.772 ± 0.047	0.781 ± 0.035	0.772 ± 0.033	0.772 ± 0.028	0.762 ± 0.025	0.766 ± 0.020
claude-3-5-haiku$_I$	0.796 ± 0.058	0.793 ± 0.033	0.786 ± 0.034	0.772 ± 0.031	0.784 ± 0.024	0.776 ± 0.021
claude-3-7-sonnet	0.790 ± 0.054	0.793 ± 0.044	0.796 ± 0.043	0.780 ± 0.033	0.788 ± 0.034	0.799 ± 0.029
claude-3-7-sonnet$_I$	0.796 ± 0.051	0.794 ± 0.049	0.801 ± 0.043	0.788 ± 0.039	0.799 ± 0.027	0.779 ± 0.034
gemini-2.0-flash	0.777 ± 0.056	0.774 ± 0.032	0.776 ± 0.031	0.770 ± 0.024	0.780 ± 0.019	0.773 ± 0.025
gemini-2.0-flash$_I$	0.794 ± 0.041	0.793 ± 0.033	0.787 ± 0.026	0.790 ± 0.026	0.789 ± 0.019	0.791 ± 0.017
gemini-2.5-flash	0.783 ± 0.046	0.795 ± 0.039	0.782 ± 0.035	0.780 ± 0.027	0.784 ± 0.024	0.788 ± 0.025
gemini-2.5-flash$_I$	0.780 ± 0.051	0.787 ± 0.031	0.777 ± 0.032	0.778 ± 0.032	0.789 ± 0.018	0.784 ± 0.022
gpt-4o	0.788 ± 0.055	0.783 ± 0.035	0.789 ± 0.025	0.793 ± 0.022	0.784 ± 0.018	0.791 ± 0.013
gpt-4o$_I$	0.815 ± 0.051	0.799 ± 0.031	0.803 ± 0.027	0.808 ± 0.024	0.809 ± 0.028	0.806 ± 0.020
gpt-4omini	0.770 ± 0.046	0.767 ± 0.034	0.768 ± 0.032	0.757 ± 0.021	0.768 ± 0.019	0.765 ± 0.020
gpt-4omini$_I$	0.769 ± 0.038	0.769 ± 0.033	0.767 ± 0.026	0.764 ± 0.021	0.767 ± 0.018	0.766 ± 0.013
gpt-o3mini	0.779 ± 0.052	0.787 ± 0.043	0.792 ± 0.023	0.784 ± 0.018	0.786 ± 0.016	0.787 ± 0.017
gpt-o3mini$_I$	0.763 ± 0.048	0.752 ± 0.032	0.763 ± 0.025	0.759 ± 0.025	0.757 ± 0.016	0.756 ± 0.017
llama4-128E	0.753 ± 0.059	0.758 ± 0.046	0.754 ± 0.028	0.763 ± 0.030	0.757 ± 0.031	0.759 ± 0.026
llama4-128E$_I$	0.772 ± 0.060	0.765 ± 0.036	0.774 ± 0.030	0.776 ± 0.030	0.773 ± 0.029	0.770 ± 0.020
llama4-16E	0.792 ± 0.049	0.778 ± 0.042	0.774 ± 0.032	0.768 ± 0.024	0.785 ± 0.031	0.774 ± 0.022
llama4-16E$_I$	0.775 ± 0.061	0.786 ± 0.041	0.780 ± 0.034	0.788 ± 0.031	0.785 ± 0.031	0.779 ± 0.028
qwen-max-vl	0.805 ± 0.046	0.796 ± 0.040	0.795 ± 0.036	0.816 ± 0.030	0.798 ± 0.034	0.801 ± 0.029
qwen-max-vl$_I$	0.785 ± 0.049	0.786 ± 0.043	0.793 ± 0.036	0.789 ± 0.030	0.803 ± 0.029	0.795 ± 0.031
qwen-plus-vl	0.784 ± 0.046	0.788 ± 0.038	0.795 ± 0.025	0.795 ± 0.024	0.798 ± 0.015	0.797 ± 0.016
qwen-plus-vl$_I$	0.765 ± 0.036	0.778 ± 0.030	0.786 ± 0.029	0.785 ± 0.022	0.783 ± 0.017	0.786 ± 0.016
Driving style - Moderate						
claude-3-5-haiku	0.772 ± 0.044	0.770 ± 0.039	0.767 ± 0.030	0.774 ± 0.023	0.772 ± 0.023	0.774 ± 0.022
claude-3-5-haiku$_I$	0.770 ± 0.046	0.783 ± 0.032	0.775 ± 0.031	0.770 ± 0.029	0.778 ± 0.026	0.773 ± 0.023
claude-3-7-sonnet	0.826 ± 0.054	0.807 ± 0.030	0.813 ± 0.027	0.807 ± 0.032	0.816 ± 0.021	0.813 ± 0.020
claude-3-7-sonnet$_I$	0.814 ± 0.049	0.807 ± 0.034	0.809 ± 0.032	0.814 ± 0.032	0.802 ± 0.024	0.805 ± 0.031
gemini-2.0-flash	0.786 ± 0.049	0.795 ± 0.031	0.797 ± 0.028	0.797 ± 0.030	0.802 ± 0.024	0.793 ± 0.028
gemini-2.0-flash$_I$	0.789 ± 0.053	0.795 ± 0.044	0.787 ± 0.040	0.793 ± 0.030	0.788 ± 0.036	0.791 ± 0.026
gemini-2.5-flash	0.784 ± 0.052	0.775 ± 0.041	0.787 ± 0.030	0.788 ± 0.026	0.780 ± 0.025	0.789 ± 0.024
gemini-2.5-flash$_I$	0.782 ± 0.070	0.789 ± 0.041	0.787 ± 0.031	0.782 ± 0.032	0.780 ± 0.028	0.791 ± 0.026
gpt-4o	0.779 ± 0.048	0.795 ± 0.046	0.797 ± 0.034	0.788 ± 0.035	0.802 ± 0.025	0.800 ± 0.029
gpt-4o$_I$	0.812 ± 0.053	0.791 ± 0.038	0.800 ± 0.034	0.797 ± 0.034	0.804 ± 0.032	0.802 ± 0.033
gpt-4omini	0.771 ± 0.042	0.785 ± 0.040	0.781 ± 0.028	0.769 ± 0.032	0.774 ± 0.037	0.771 ± 0.028
gpt-4omini$_I$	0.763 ± 0.056	0.778 ± 0.039	0.782 ± 0.035	0.787 ± 0.034	0.773 ± 0.030	0.780 ± 0.030
gpt-o3mini	0.810 ± 0.044	0.793 ± 0.035	0.802 ± 0.029	0.805 ± 0.029	0.803 ± 0.024	0.806 ± 0.020
gpt-o3mini$_I$	0.773 ± 0.047	0.788 ± 0.035	0.766 ± 0.028	0.773 ± 0.030	0.775 ± 0.028	0.769 ± 0.025
llama4-128E	0.788 ± 0.062	0.780 ± 0.035	0.777 ± 0.035	0.775 ± 0.029	0.776 ± 0.028	0.776 ± 0.025
llama4-128E$_I$	0.792 ± 0.036	0.793 ± 0.032	0.795 ± 0.028	0.787 ± 0.019	0.790 ± 0.014	0.789 ± 0.016
llama4-16E	0.813 ± 0.041	0.784 ± 0.037	0.799 ± 0.038	0.801 ± 0.035	0.802 ± 0.035	0.807 ± 0.030
llama4-16E$_I$	0.797 ± 0.038	0.821 ± 0.032	0.808 ± 0.029	0.811 ± 0.027	0.809 ± 0.024	0.813 ± 0.023
qwen-max-vl	0.787 ± 0.042	0.779 ± 0.051	0.779 ± 0.041	0.792 ± 0.028	0.791 ± 0.037	0.791 ± 0.031
qwen-max-vl$_I$	0.806 ± 0.053	0.773 ± 0.048	0.760 ± 0.037	0.797 ± 0.036	0.784 ± 0.037	0.793 ± 0.035
qwen-plus-vl	0.792 ± 0.044	0.791 ± 0.036	0.791 ± 0.032	0.797 ± 0.028	0.792 ± 0.025	0.791 ± 0.025
qwen-plus-vl$_I$	0.755 ± 0.057	0.766 ± 0.041	0.766 ± 0.037	0.765 ± 0.032	0.762 ± 0.033	0.760 ± 0.034

Table A.5
Variance of B_{sim} across different sequence length L for the *driver gender* category. Results are reported as mean ± standard deviation. The subscript I denotes models that use images as input.

Model	$L = 10$	$L = 20$	$L = 30$	$L = 40$	$L = 50$	$L = 60$
Driving years - <=5 years						
claude-3-5-haiku	0.779 ± 0.055	0.784 ± 0.035	0.777 ± 0.034	0.770 ± 0.029	0.773 ± 0.025	0.769 ± 0.028
claude-3-5-haiku$_I$	0.790 ± 0.058	0.786 ± 0.027	0.776 ± 0.038	0.781 ± 0.020	0.773 ± 0.026	0.771 ± 0.023
claude-3-7-sonnet	0.797 ± 0.048	0.816 ± 0.034	0.812 ± 0.031	0.798 ± 0.031	0.799 ± 0.026	0.798 ± 0.026
claude-3-7-sonnet$_I$	0.816 ± 0.049	0.798 ± 0.048	0.809 ± 0.035	0.802 ± 0.031	0.811 ± 0.032	0.802 ± 0.030
gemini-2.0-flash	0.789 ± 0.046	0.801 ± 0.044	0.789 ± 0.043	0.792 ± 0.038	0.799 ± 0.027	0.782 ± 0.034
gemini-2.0-flash$_I$	0.803 ± 0.044	0.799 ± 0.044	0.786 ± 0.042	0.794 ± 0.046	0.783 ± 0.035	0.802 ± 0.028
gemini-2.5-flash	0.784 ± 0.054	0.780 ± 0.047	0.789 ± 0.038	0.775 ± 0.036	0.783 ± 0.030	0.794 ± 0.027
gemini-2.5-flash$_I$	0.773 ± 0.050	0.778 ± 0.040	0.801 ± 0.039	0.782 ± 0.037	0.793 ± 0.028	0.776 ± 0.040
gpt-4o	0.803 ± 0.057	0.794 ± 0.033	0.804 ± 0.031	0.793 ± 0.035	0.788 ± 0.031	0.794 ± 0.035
gpt-4o$_I$	0.814 ± 0.060	0.798 ± 0.045	0.809 ± 0.036	0.801 ± 0.036	0.798 ± 0.035	0.808 ± 0.029
gpt-4omini	0.777 ± 0.061	0.781 ± 0.036	0.770 ± 0.042	0.782 ± 0.037	0.780 ± 0.034	0.780 ± 0.027
gpt-4omini$_I$	0.774 ± 0.048	0.783 ± 0.040	0.773 ± 0.036	0.777 ± 0.036	0.773 ± 0.032	0.783 ± 0.025
gpt-o3mini	0.801 ± 0.052	0.793 ± 0.042	0.801 ± 0.035	0.798 ± 0.031	0.804 ± 0.036	0.798 ± 0.030
gpt-o3mini$_I$	0.751 ± 0.056	0.768 ± 0.040	0.773 ± 0.031	0.776 ± 0.032	0.764 ± 0.037	0.770 ± 0.028
llama4-128E	0.778 ± 0.042	0.773 ± 0.053	0.774 ± 0.038	0.765 ± 0.039	0.771 ± 0.036	0.777 ± 0.035
llama4-128E$_I$	0.797 ± 0.048	0.782 ± 0.037	0.797 ± 0.035	0.794 ± 0.034	0.783 ± 0.038	0.799 ± 0.026
llama4-16E	0.796 ± 0.051	0.808 ± 0.039	0.800 ± 0.038	0.804 ± 0.037	0.796 ± 0.029	0.798 ± 0.034
llama4-16E$_I$	0.797 ± 0.048	0.809 ± 0.047	0.812 ± 0.042	0.806 ± 0.038	0.803 ± 0.035	0.814 ± 0.022
qwen-max-vl	0.796 ± 0.062	0.809 ± 0.048	0.810 ± 0.040	0.803 ± 0.040	0.803 ± 0.038	0.806 ± 0.035
qwen-max-vl$_I$	0.799 ± 0.063	0.792 ± 0.050	0.793 ± 0.048	0.793 ± 0.041	0.796 ± 0.039	0.801 ± 0.037
qwen-plus-vl	0.814 ± 0.035	0.797 ± 0.034	0.799 ± 0.031	0.795 ± 0.034	0.799 ± 0.026	0.795 ± 0.025
qwen-plus-vl$_I$	0.775 ± 0.043	0.771 ± 0.041	0.784 ± 0.039	0.783 ± 0.028	0.772 ± 0.034	0.777 ± 0.029
Driving years - >5 years						
claude-3-5-haiku	0.783 ± 0.047	0.774 ± 0.042	0.777 ± 0.035	0.789 ± 0.029	0.781 ± 0.028	0.780 ± 0.028
claude-3-5-haiku$_I$	0.791 ± 0.051	0.786 ± 0.027	0.781 ± 0.034	0.789 ± 0.030	0.790 ± 0.024	0.793 ± 0.027
claude-3-7-sonnet	0.798 ± 0.054	0.815 ± 0.032	0.822 ± 0.032	0.815 ± 0.029	0.818 ± 0.024	0.815 ± 0.024
claude-3-7-sonnet$_I$	0.806 ± 0.057	0.822 ± 0.047	0.820 ± 0.037	0.822 ± 0.034	0.821 ± 0.035	0.814 ± 0.030
gemini-2.0-flash	0.797 ± 0.040	0.803 ± 0.038	0.811 ± 0.030	0.801 ± 0.027	0.808 ± 0.022	0.805 ± 0.023
gemini-2.0-flash$_I$	0.785 ± 0.038	0.802 ± 0.032	0.796 ± 0.024	0.799 ± 0.024	0.795 ± 0.022	0.793 ± 0.016
gemini-2.5-flash	0.803 ± 0.043	0.783 ± 0.028	0.793 ± 0.038	0.798 ± 0.030	0.786 ± 0.025	0.800 ± 0.021
gemini-2.5-flash$_I$	0.799 ± 0.047	0.807 ± 0.029	0.795 ± 0.026	0.795 ± 0.027	0.801 ± 0.020	0.803 ± 0.022
gpt-4o	0.804 ± 0.049	0.806 ± 0.030	0.809 ± 0.028	0.795 ± 0.027	0.811 ± 0.022	0.814 ± 0.022
gpt-4o$_I$	0.817 ± 0.038	0.819 ± 0.037	0.814 ± 0.028	0.812 ± 0.031	0.814 ± 0.030	0.813 ± 0.028
gpt-4omini	0.781 ± 0.052	0.782 ± 0.041	0.779 ± 0.039	0.769 ± 0.031	0.778 ± 0.033	0.771 ± 0.028
gpt-4omini$_I$	0.784 ± 0.062	0.776 ± 0.031	0.774 ± 0.031	0.776 ± 0.028	0.777 ± 0.024	0.769 ± 0.025
gpt-o3mini	0.807 ± 0.049	0.809 ± 0.030	0.807 ± 0.030	0.810 ± 0.023	0.810 ± 0.023	0.806 ± 0.020
gpt-o3mini$_I$	0.776 ± 0.055	0.773 ± 0.032	0.772 ± 0.028	0.773 ± 0.031	0.766 ± 0.030	0.768 ± 0.025
llama4-128E	0.794 ± 0.046	0.786 ± 0.035	0.786 ± 0.036	0.782 ± 0.032	0.786 ± 0.024	0.779 ± 0.026
llama4-128E$_I$	0.785 ± 0.059	0.800 ± 0.037	0.802 ± 0.035	0.807 ± 0.035	0.801 ± 0.033	0.803 ± 0.024
llama4-16E	0.805 ± 0.055	0.789 ± 0.039	0.802 ± 0.032	0.804 ± 0.029	0.801 ± 0.027	0.807 ± 0.027
llama4-16E$_I$	0.815 ± 0.044	0.812 ± 0.036	0.817 ± 0.034	0.821 ± 0.028	0.805 ± 0.032	0.804 ± 0.032
qwen-max-vl	0.807 ± 0.049	0.803 ± 0.050	0.810 ± 0.040	0.791 ± 0.039	0.792 ± 0.039	0.795 ± 0.040
qwen-max-vl$_I$	0.802 ± 0.053	0.795 ± 0.047	0.791 ± 0.045	0.800 ± 0.041	0.792 ± 0.034	0.790 ± 0.036
qwen-plus-vl	0.778 ± 0.043	0.795 ± 0.035	0.798 ± 0.028	0.800 ± 0.030	0.789 ± 0.028	0.789 ± 0.026
qwen-plus-vl$_I$	0.749 ± 0.052	0.754 ± 0.031	0.760 ± 0.032	0.760 ± 0.036	0.754 ± 0.032	0.763 ± 0.028

References

Anthropic, 2024. The claude 3 model family: opus, sonnet, haiku. https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T., 2023. Qwen technical report. arXiv preprint arXiv: 2309.16609. https://arxiv.org/abs/2309.16609.

Barberis, N., Thaler, R., 2003. A survey of behavioral finance. Handbook Econ. Finance 1, 1053–1128.

Bhattacharyya, R., Wulfe, B., Phillips, D.J., Kuefler, A., Morton, J., Senanayake, R., Kochenderfer, M.J., 2023. Modeling human driving behavior through generative adversarial imitation learning. IEEE Trans. Intell. Transp. Syst. 24 (3), 2874-2887. https://doi.org/10.1109/tits.2022.3227738

Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C.V., Karatzas, D., 2019. Scene text visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 4291–4301.

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuscenes: a multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11621–11631.

Cai, R., Li, B., Wen, X., Chen, M., Zhao, Z., 2025. Diagnosing and mitigating modality interference in multimodal large language models. arXiv preprint arXiv: 2505.19616. https://arxiv.org/abs/2505.19616.

Chang, T.A., Rajagopal, D., Bolukbasi, T., Dixon, L., Tenney, I., 2024. Scalable influence and fact tracing for large language model pretraining. arXiv preprint arXiv:2410.17413 .

Chen, H., Covert, I.C., Lundberg, S.M., Lee, S.-I., 2023. Algorithms to estimate shapley value feature attributions. Nature Mach. Intell. 5 (6), 590–601.

Chen, J., Lu, S., 2024. An advanced driving agent with the multimodal large language model for autonomous vehicles. In: 2024 IEEE International Conference on Mobility, Operations, Services and Technologies (MOST). IEEE, pp. 1–11.

Chen, M., Cao, Y., Zhang, Y., Lu, C., 2024a. Quantifying and mitigating unimodal biases in multimodal large language models: a causal perspective. arXiv preprint arXiv: 2403.18346. https://arxiv.org/abs/2403.18346.

Chen, Q., Zhao, D., Liu, C., Yang, M., Shi, Y., 2024b. Autonomous vehicles in mixed-autonomy traffic: game theoretic human-like decision making countermeasures. Complex Eng. Syst. 4 (4), N–A.

Cui, Y., Huang, S., Zhong, J., Liu, Z., Wang, Y., Sun, C., Li, B., Wang, X., Khajepour, A., 2023. Drivellm: Charting the path toward full autonomous driving with large language models. IEEE Trans. Intell. Veh. 9 (1), 1450–1464.

Deo, N., Trivedi, M.M., 2018. Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms. In: 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, pp. 1179–1184.

Ding, W., Chen, J., Shen, S., 2019. Predicting vehicle behaviors over an extended horizon using behavior interaction network. In: 2019 International Conference on Robotics and Automation (ICRA). IEEE, pp. 8634–8640.

Gao, S., Wu, X., Dimakis, A., Courcoubetis, C., 2024. The impact of autonomous vehicles on ride-hailing platforms with strategic human drivers. arXiv preprint arXiv:2406.19014 .

Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., et al., 2020. A2d2: Audi autonomous driving dataset. arXiv preprint arXiv:2004.06320 .

Goldshmidt, R., Horovicz, M., 2024. TokenSHAP: interpreting large language models with monte carlo shapley value estimation. arXiv preprint arXiv:2407.10114 .

Goutte, C., Gaussier, E., 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: European Conference on Information Retrieval. Springer, pp. 345–359.

Han, W., Guo, D., Xu, C.-Z., Shen, J., 2025. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39, pp. 3347–3355.

Hang, P., Lv, C., Xing, Y., Huang, C., Hu, Z., 2020. Human-like decision making for autonomous driving: A noncooperative game theoretic approach. IEEE Trans. Intell. Transp. Syst. 22 (4), 2076–2087.

Hou, X., Wang, W., Yang, L., Lin, H., Feng, J., Min, H., Zhao, X., 2025. Driveagent: multi-agent structured reasoning with LLM and multimodal sensor fusion for autonomous driving. IEEE Rob. Automation Lett. 10 (11), 12189–12196. https://doi.org/10.1109/LRA.2025.3619807

Hu, B., Ray, B., Leung, A., Summerville, A., Joy, D., Funk, C., Basharat, A., 2024. Language models are alignable decision-makers: dataset and application to the medical triage domain. arXiv preprint arXiv:2406.06435 .

Hu, X., Zheng, Z., Chen, D., Zhang, X., Sun, J., 2022. Processing, assessing, and enhancing the waymo autonomous vehicle open dataset for driving behavior research. Transp. Res.. Part C: Emerg. Technol. 134, 103490.

Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A.J., Welihinda, A., Hayes, A., Radford, A., et al., 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276 .

Islam, M.R., 2024. Application of multimodal large language models in autonomous driving. arXiv preprint arXiv:2412.16410 .

Janai, J., Güney, F., Behl, A., Geiger, A., et al., 2020. Computer vision for autonomous vehicles: problems, datasets and state of the art. Found. trends® in computer graphics and vision 12 (1–3), 1–308.

Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J., 2024. Bench2drive: towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. arXiv preprint arXiv:2406.03877 .

Jin, Y., Shen, X., Peng, H., Liu, X., Qin, J., Li, J., Xie, J., Gao, P., Zhou, G., Gong, J., 2023. Surrealdriver: designing generative driver agent simulation framework in urban contexts based on large language model. arXiv preprint arXiv:2309.13193 5 (7), 8.

Jin, Y., Yang, R., Yi, Z., Shen, X., Peng, H., Liu, X., Qin, J., Li, J., Xie, J., Gao, P., et al., 2024. Surrealdriver: designing LLM-powered generative driver agent framework based on human drivers' driving-thinking data. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 966–971.

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30 (1-2), 81–93.

Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S., Robnik-Šikonja, M., 2021. BERT meets shapley: Extending SHAP explanations to transformer-based classifiers. In: Toivonen, H., Boggia, M. (Eds.), Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. Association for Computational Linguistics, Online, pp. 16–21. https://aclanthology.org/2021.hackashop-1.3.

Koorathota, S., Papadopoulos, N., Ma, J.L., Kumar, S., Sun, X., Mittal, A., Adelman, P., Sajda, P., 2024. Gaze-informed vision transformers: predicting driving decisions under uncertainty. In: Companion Proceedings of the 26th International Conference on Multimodal Interaction, pp. 184–194.

Kujanpää, K., Baimukashev, D., Zhu, S., Azam, S., Munir, F., Alcan, G., Kyrki, V., 2024. Challenges of data-driven simulation of diverse and consistent human driving behaviors. arXiv preprint arXiv:2401.03236 .

Lee, S., Wang, Z.J., Chakravarthy, A., Helbling, A., Peng, S., Phute, M., Chau, D. H.P., Kahng, M., 2025. Llm attributor: Interactive visual attribution for llm generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39, pp. 29655–29657.

Lee, Y., Lee, S., Kim, S., . Csg-driver: common sense guided autonomous driving under legal compliance and practical flexibility in dilemma situations. In: AAAI 2025 Workshop LM4Plan.

Levenshtein, V.I., et al., 1966. Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. Vol. 10. Soviet Union, pp. 707–710.

Li, S., 2024. AI-Empowered Personalized Prediction and Decision-Making Systems for Driving Co-Pilot. University of California, Riverside.

Liu, L., Wang, P., Wu, G., Jiang, J., Yang, H., 2026. Towards optimal mixture of experts system for 3d object detection: A game of accuracy, efficiency and adaptivity. IEEE Trans. Pattern Anal. Mach. Intell. 48 (1), 914–931. https://doi.org/10.1109/TPAMI.2025.3611795

Liu, Z., Sun, Y., Wang, Y., Yang, N., Li, S.E., Zhao, X., 2025b. Beyond conventional vision: RGB-event fusion for robust object detection in dynamic traffic scenarios. 5, 100202. arXiv preprint arXiv: 2508.10704, https://doi.org/10.1016/j.commtr.2025.100202.

Long, K., Shi, H., Liu, J., Li, X., 2024. Vlm-mpc: vision language foundation model (vlm)-guided model predictive controller (mpc) for autonomous driving. arXiv preprint arXiv:2408.04821 .
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.
Ma, Y., Cui, C., Cao, X., Ye, W., Liu, P., Lu, J., Abdelraouf, A., Gupta, R., Han, K., Bera, A., Rehg, J.M., Wang, Z., 2024. LaMPilot: an open benchmark dataset for autonomous driving with language model programs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15141–15151.
Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., et al., 2021. One million scenes for autonomous driving: Once dataset. arXiv preprint arXiv:2106.11037 .
Mao, J., Qian, Y., Ye, J., Zhao, H., Wang, Y., 2023. Gpt-driver: learning to drive with gpt. arXiv preprint arXiv:2310.01415 .
Musker, S., Duchnowski, A., Millière, R., Pavlick, E., 2024. Semantic structure-mapping in LLM and human analogical reasoning. arXiv preprint arXiv:2406.13803 .
Nam, D., Macvean, A., Hellendoorn, V., Vasilescu, B., Myers, B., 2024. Using an llm to help with code understanding. In: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, pp. 1–13.
Orrell, D., 2010. Economyths: Ten Ways Economics Gets it Wrong. John Wiley & Sons.
Ou, J., Lu, J., Liu, C., Tang, Y., Zhang, F., Zhang, D., Gai, K., 2024. Dialogbench: evaluating LLMs as human-like dialogue systems. arXiv preprint arXiv: 2311.01677. https://arxiv.org/abs/2311.01677.
Paek, D.-H., Kong, S.-H., Wijaya, K.T., 2022. K-radar: 4d radar object detection for autonomous driving in various weather conditions. Adv. Neural Inf. Process. Syst. 35, 3819–3829.
Park, S., Lee, M., Kang, J., Choi, H., Park, Y., Cho, J., Lee, A., Kim, D., 2024. Vlaad: vision and language assistant for autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 980–987.
Rock, T., Hryhoruk, T., Bleher, T., Bahram, M., Marker, S., Syed, A.A., Sekeran, M., 2023. Objectively scoring the human-likeness of artificial driver models. Appl. Sci. 13 (18), 10218.
Sha, H., Mu, Y., Jiang, Y., Zhan, G., Chen, L., Xu, C., Luo, P., Li, S.E., Tomizuka, M., Zhan, W., Ding, M.,et al., 2025. LanguageMPC:Large language models as decision makers for autonomous driving. arXiv:2310.03026
Shapley, L.S., 1953. A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (Eds.), Contributions to the Theory of Games. Princeton University Press. Vol. 2, pp. 307–317.
Shen, Z., 2024. Llm with tools: a survey. arXiv preprint arXiv:2409.18807 .
Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H., 2024. Drivelm: Driving with graph visual question answering. In: European Conference on Computer Vision. Springer, pp. 256–274.
Singh, A., 2025. Meta llama 4: the future of multimodal AI. SSRN Electron. J. , 15. https://doi.org/10.2139/ssrn.5208228
Sokolova, M., Japkowicz, N., Szpakowicz, S., 2006. Beyond accuracy, f-score and ROC: a family of discriminant measures for performance evaluation. In: Australasian joint conference on artificial intelligence. Springer, pp. 1015–1021.
Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manag. 45 (4), 427–437.
Stead, A., Michael, P., Ondo, K., 2022. Student navigation through computer-based simulations: What predicts success? Teach. Learn. Commun. Sci. Disorders 6 (1), 10.
Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al., 2020. Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2446–2454.
Sun, Y., Salami Pargoo, N., Jin, P., Ortiz, J., 2024. Optimizing autonomous driving for safety: a human-centric approach with llm-enhanced rlhf. In: Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 76–80.
Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al., 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 .
Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 .
Vinayavekhin, P., Tatsubori, M., Kimura, D., Huang, Y., De Magistris, G., Munawar, A., Tachibana, R., 2017. Human-like hand reaching by motion prediction using long short-term memory. In: Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9. Springer, pp. 156–166.
Wang, W., Xi, J., Chen, H., 2014. Modeling and recognizing driver behavior based on driving data: A survey. Math. Problems Eng. 2014 (1), 245641.
Wang, Y., Jiao, R., Zhan, S.S., Lang, C., Huang, C., Wang, Z., Yang, Z., Zhu, Q., 2023. Empowering autonomous driving with large language models: A safety perspective. arXiv preprint arXiv:2312.00812 .
Wu, D., Han, W., Liu, Y., Wang, T., Xu, C.-z., Zhang, X., Shen, J., 2025a. Language prompt for autonomous driving. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 39, pp. 8359–8367.
Wu, G., Su, B., Zhao, Y., Wang, P., Lin, Y., Yang, H.F., 2025b. Towards physics-informed spatial intelligence with human priors: An autonomous driving pilot study. arXiv preprint arXiv:2510.21160 .
Wu, K., Li, W., Xiao, X., 2024a. AccidentGPT: large multi-modal foundation model for traffic accident analysis. arXiv preprint arXiv:2401.03040 .
Wu, S., Oltramari, A., Francis, J., Giles, C.L., Ritter, F.E., 2024b. Cognitive LLMs: Toward human-like artificial intelligence by integrating cognitive architectures and large language models for manufacturing decision-making. Neurosymbolic Artif. Intell. .
Yang, R., Zhang, X., Fernandez-Laaksonen, A., Ding, X., Gong, J., 2024. Driving style alignment for llm-powered driver agent. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 11318–11324.
Yang, Z., Jia, X., Li, H., Yan, J., 2023a. Llm4drive: a survey of large language models for autonomous driving. arXiv preprint arXiv:2311.01043 .
Yang, Z., Jia, X., Li, H., Yan, J., 2023b. Llm4drive: a survey of large language models for autonomous driving. arXiv preprint arXiv: 2311.01043.
Yang, Z., Yang, J., Zhou, Y., Xu, Q., Ou, M., 2025. Multimodal trajectory prediction for intelligent connected vehicles in complex road scenarios based on causal reasoning and driving cognition characteristics. Sci. Rep. 15 (1), 7259.
You, J., Jia, X., Zhang, Z., Zhu, Y., Yan, J., 2024. Bench2drive-r: turning real world data into reactive closed-loop autonomous driving benchmark by generative model. arXiv preprint arXiv:2412.09647 .
Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T., 2020. Bdd100k: a diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2636–2645.
Zeng, W., Luo, W., Suo, S., Sadat, A., Yang, B., Casas, S., Urtasun, R., 2019. End-to-end interpretable neural motion planner. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8660–8669.
Zhang, P., Cheng, T., Jiang, Y., Liu, W., Chen, X., Zou, X., Sheng, Y., Zhao, X., 2025a. Advancements and insights in assessing cognitive load during driving: a comprehensive narrative review. IEEE Trans. Intell. Transp. Syst. PP, 1–17. https://doi.org/10.1109/TITS.2025.3640416
Zhang, R., Wang, B., Zhang, J., Bian, Z., Feng, C., Ozbay, K., 2025b. When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis. arXiv preprint arXiv:2501.10604 .
Zhang, R., Yang, Z., Dai, T., Jian Sun, D., Ma, S., Zhao, X., 2025c. Magae: Multi-level alignment over aggregation semantic graph with attribute enhancement for text-based vehicle retrieval. IEEE Trans. Intell. Transp. Syst. 26 (9), 13704–13720. https://doi.org/10.1109/TITS.2025.3566502
Zhao, Y., Wang, P., Zhao, Y., Du, H., Yang, H.F., 2025a. Safetraffic copilot: adapting large language models for trustworthy traffic safety assessments and decision interventions. Nature Commun. 16 (1), 8846.
Zhao, Y., Zhao, Y., Du, H., Yang, H.F., 2025b. Personalized decision modeling: Utility optimization or textualized-symbolic reasoning. arXiv preprint arXiv:2511.02194.

Zhou, L., Jiang, M., Wang, D., 2024. Humansim: human-like multi-agent novel driving simulation for corner case generation. In: ECCV 2024 Workshop on Multimodal Perception and Comprehension of Corner Cases in Autonomous Driving.
Zhu, M., Wang, X., Wang, Y., 2018. Human-like autonomous car-following model with deep reinforcement learning. Transp. Res. Part C: Emerg. Technol. 97, 348–368.