Transportation Research Part C 180 (2025) 105361

Contents lists available at ScienceDirect

Transportation Research Part C

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

Attribution explanations for decision-making in deep lane-change models

Rui Shi [a], Tianxing Li [b,*], Yasushi Yamaguchi [c], Liguo Zhang [a]

[a] School of Information Science and Technology, Beijing University of Technology, Beijing, 100124, China
[b] College of Computer Science, Beijing University of Technology, Beijing, 100124, China
[c] Department of General Systems Studies, University of Tokyo, Tokyo, 153-8902, Japan

ARTICLE INFO

Keywords:
Autonomous driving
Lane change
Attribution explanation
Explainable artificial intelligence

ABSTRACT

Deep learning models are attracting considerable attention for their potential to enable intelligent lane-change behaviors. To ensure the reliability of these models, it is essential to understand their decision-making processes using attribution methods. However, existing attribution techniques, which are predominantly developed for visual tasks, often struggle to deliver accurate and interpretable explanations when applied to complex lane-change models. We identify the essential cause of this problem as the high variability in the distribution of connected perceptual information that serves as input to lane-change models. To address this, we propose a novel path aggregation attribution method, where paths describe the transition of a feature from absence to presence, expressing its relative contribution. Our method leverages an exponential family to introduce probabilistic paths and calculate attribution expectations, effectively traversing the input feature distribution space to provide a more comprehensive representation of feature transitions. Additionally, we introduce a distribution-informed counterfactual reference to define starting points of the paths, enabling the flexible generation of traffic scenarios with feature absence. Extensive experiments on three lane-change models show that our method consistently outperforms state-of-the-art attribution methods. Specifically, we achieve higher performance on four widely used quantitative metrics, *i.e.*, sensitivity-n, accuracy information curve, softmax information curve, and most-relevant-first, demonstrating superior reliability and interpretability.

1. Introduction

Deep learning has experienced rapid advancements and is increasingly being employed in critical transportation applications (Dong et al., 2023), such as driving behavior prediction (Anik et al., 2024; Sun and Yang, 2024), trajectory tracking (Wang et al., 2024; Xue et al., 2024), localization and mapping (Liang et al., 2023; Li et al., 2025a), and path planning (Du et al., 2023; Su et al., 2024). Among these, autonomous lane-change (Ali et al., 2025) is one of the most challenging yet essential tasks for intelligent transport systems. Given the high-risk nature of lane-change operations, understanding how black-box models generate their outputs is essential to ensure reliable deployment and to enhance safety.

Although numerous attribution explanation studies, predominantly developed for image classification tasks (Shi et al., 2025b), have emerged to address this need, the complex feature distribution in lane-change scenarios presents unique challenges. First, unlike vision tasks where the input consists of a uniform matrix of pixels, lane-change models process network-connected perceptual

* Corresponding author.
E-mail address: litianxing@bjut.edu.cn (T. Li).

https://doi.org/10.1016/j.trc.2025.105361
Received 28 March 2025; Received in revised form 15 September 2025; Accepted 23 September 2025

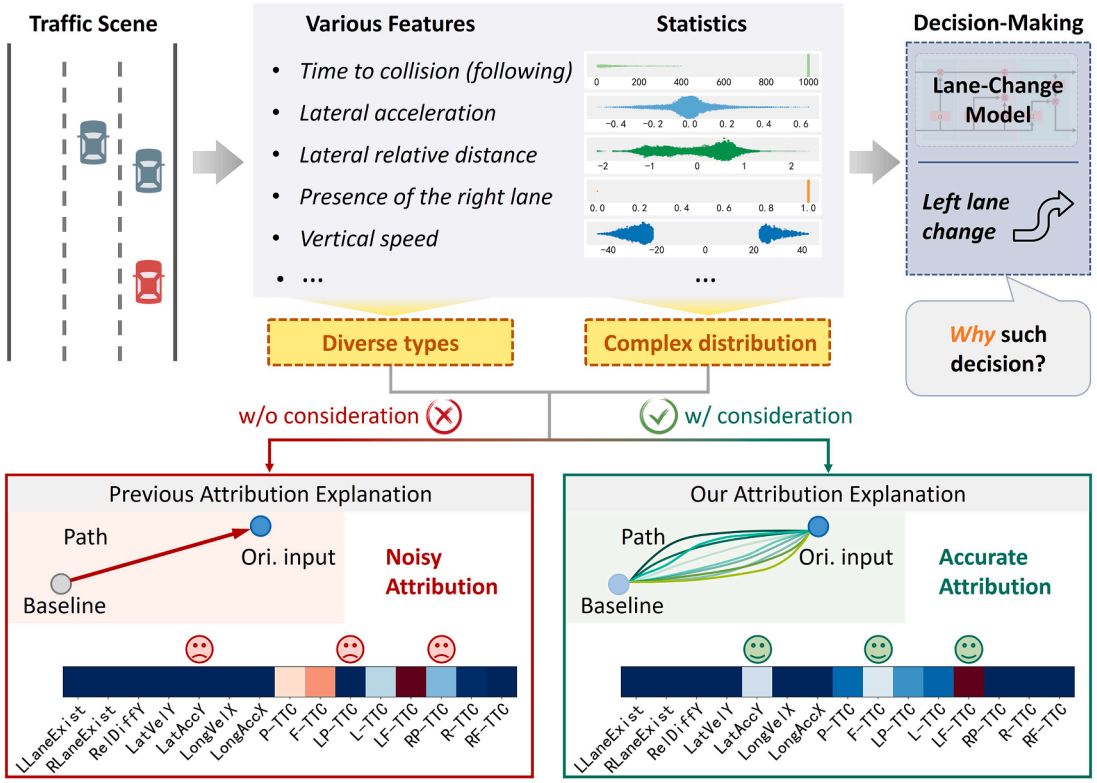

Fig. 1. Conceptual illustration of our idea for generating attribution explanation.

data, such as distances to surrounding vehicles, time-to-collision, and their corresponding speeds and accelerations (Gupta et al., 2024; Zhan et al., 2025). The diverse nature of these features makes it difficult for attribution methods to accurately interpret the underlying reasoning process. Moreover, the distributions of these inputs vary significantly, with such variability directly reflected in their numerical values. For instance, when no preceding vehicle is present, many models define the relative distance as infinity, while speed and acceleration values typically remain within narrower and more specific ranges. A comparison of these data distributions, visualized using the highD dataset (Krajewski et al., 2018) and shown in the upper middle part of Fig. 1, highlights the substantial variability in feature distributions.

Path-based attribution methods, such as the Aumann-Shapley method (Aumann and Shapley, 1974), provide a promising framework for addressing the challenges of feature attribution by describing the transition from feature absence to presence. These methods compute the relative contributions of input features along predefined paths representing this transition. However, traditional approaches (Aumann and Shapley, 1974; Shi et al., 2024), which often rely on a single straight-line path between a reference and the input data, struggle to capture the complexities of real-world feature distributions, particularly in deep lane-change models. This limitation can introduce out-of-distribution artifacts, resulting in noisy and inaccurate attributions, such as incorrectly associating a left lane-change decision with the time-to-collision of the right preceding vehicle (RP-TTC) when no right lane exists, as shown in Fig. 1. To overcome this, we propose a probabilistic path-based method that leverages an exponential family of paths to better traverse the feature distribution space. By statistically mitigating errors caused by specific path definitions and aggregating attributions across probabilistic paths, our method ensures that input features adequately traverse their effective distributions during transitions. This statistically robust approach prioritizes accuracy over deterministic precision, enabling more reliable attributions, such as correctly identifying lateral acceleration (LatAccY) as a key feature in left lane-change decisions, as shown in Fig. 1.

From a technical implementation perspective, we use the Dirichlet distribution to efficiently sample from the vast space of possible attribution paths. The concentration parameter of this distribution is dynamically adjusted according to the sparsity of input features: sparser distributions receive lower concentration values, while denser distributions receive higher values. During our methodology development, we evaluated several probability distributions for path definition. Although several members of the exponential family of distributions demonstrated satisfactory performance, the Dirichlet distribution proved superior due to its flexible concentration parameter and its consistently strong empirical performance across test scenarios.

Additionally, we introduce an innovative counterfactual generation method to determine the reference point, *i.e.*, the starting point of an attribution path. The reference is designed to represent feature absence while maintaining proximity to the actual data distribution, enabling more reliable attribution calculations.

To sum up, our main contributions are as follows:

- To address the challenges posed by the complex feature distributions of lane-change models, we redefine the Aumann-Shapley path by introducing probabilistic paths generated using the Dirichlet distribution. This approach effectively mitigates attribution inaccuracy caused by out-of-distribution issues.
- We develop an optimization-based counterfactual method for determining a reference point that accurately represents the absence of features while respecting the underlying data distribution. This optimized reference point leads to the generation of more accurate and reliable relative contributions (*i.e.*, attributions).

We validate our method using three representative lane-change models: a time-series model based on long short-term memory (LSTM) (Wang et al., 2022), a convolutional neural network (CNN) model (Zhang et al., 2023b), and an attention-based Transformer model (Gao et al., 2023), all demonstrating strong performance. Furthermore, to demonstrate the practical utility of our attribution method in model analysis, we design two experiments: one evaluates feature statistical contribution, and the other analyzes the alignment between human and model decisions.

The remainder of the paper is organized as follows. Section 2 introduces attribution methods and deep lane change models. Section 3 describes the proposed attribution method. Section 4 presents the experimental results, including ablation studies, comparisons with state-of-the-art attributionmethods, and lane-change model analysis based on attribution explanations. Finally, Section 5 concludes the paper.

2. Related work

2.1. Feature attribution methods

Feature attribution methods provide essential insights into the decision-making process of deep neural networks (DNNs). These methods can be categorized as local post-hoc explanation techniques. In contrast to global methods, attribution explanations emphasize specific decision instances rather than providing a comprehensive overview of the model's overall behavior. They assign importance scores to input features, revealing each feature's contribution to the model's output (Dong et al., 2023; Zhang and Li, 2022). While earlier attribution methods often manipulated inputs, recent advancements in graphics processing units (GPUs) and deep learning libraries have enabled efficient forward and backward propagation, leading to a preference for gradient-based attribution methods. Our proposed explanation method, grounded in Aumann-Shapley (AS) values, also leverages these computational advantages.

A prominent line of work focuses on perturbing input features and observing the resulting changes in network output. SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) for the interpretation of machine learning models is a widely adopted perturbation-based method. It employs the Shapley value from cooperative game theory to fairly distribute the output among input features. Li et al. (2024b) adapted the SHAP paradigm to propose SVCE for interpreting lane-change decisions, a method that can be readily extended to other intelligent transportation models. By designing counterfactual samples, their approach could, to a certain extent, address the issue of ambiguous baseline definitions present in the original SHAP. More recently, Li et al. (2024d) introduced a fast Shapley value estimation method, combining the results with activation maps to generate GT-CAM attribution explanations. Although such perturbation-based frameworks offer remarkable flexibility and broad applicability, they can be computationally expensive, as each input modification necessitates a new model evaluation.

Leveraging increasingly efficient DNN feature propagation, researchers have focused on gradient-based techniques. Yang et al. (2023) introduced important direction gradient integration (IDGI), which identifies gradient directions that minimize noise in pixel-level attribution. Li et al. (2024a) proposed the expected integral discrete gradient to interpret various machine learning models for lane-change prediction. Their method equips users with essential tools to better understand and analyze model behavior. Zhang et al. (2024) developed SAMP, which identifies near-optimal manipulation paths to accumulate gradients for attribution. Chormai et al. (2024) refined layer-wise relevance propagation (LRP) by incorporating principal and independent component analyzes to highlight the most active components within the DNN. Furthermore, Li et al. (2025d) investigated the roles of weights and sample features in LRP, leading to the development of weight-dependent baseline LRP (WB-LRP) for graph convolutional neural networks. GradShap (Lundberg and Lee, 2017) also remains a popular tool, frequently applied in real-world driving scenarios like object detection (Oseni et al., 2023) and lane-change prediction (Li et al., 2023b). Shi et al. (2025a) introduced multiple integration to improve AS attribution calculation and implemented visualization. Methods originating from fields such as biology, economics, chemical engineering, geometry, and psychology (Chen et al., 2023; Janizek et al., 2023; Li et al., 2024c; Kim et al., 2024; Li et al., 2025b) also offer valuable insights for enhancing explainability in intelligent transportation models. Although these methods possess some generalizability, they often do not account for differences in input feature distributions. This oversight can be problematic for deep lane-change models, where significant distribution differences can substantially reduce the accuracy of attribution calculations.

Further research explores combining multiple attribution techniques, such as refining LRP through optimization (Bassi et al., 2024) or integrating diverse attribution perspectives (Kapishnikov et al., 2019) to better illustrate the influence of input features on model output. However, many existing solutions are highly specialized, hindering their broader adaptation and generalization across different driving models. This specialization creates challenges for consistent application and can impact the reliability and comparability of the resulting explanations.

While previous attribution methods such as LRP and GradShap have achieved notable success, they often overlook the impact of complex or sparse input distributions typical of lane-change scenarios, which can lead to unreliable explanations. In contrast, our

approach leverages Dirichlet-distributed probabilistic paths and an optimized counterfactual reference, specifically designed to adapt to real-world data sparsity and distribution shifts. This makes our attributions more robust and accurate, addressing a key limitation of existing techniques.

2.2. State-of-the-art lane-change models

In recent years, significant progress has been made in lane-change modeling, including both deep learning-based models and non-deep learning methods. Deep learning models often employ either sequential models like recurrent neural networks, particularly gated recurrent unit (GRU) and long short-term memory (LSTM), convolutional neural networks (CNNs), or Transformer networks (Xing et al., 2025; Yao and Sun, 2025). Sequential models leverage LSTMs to process temporal data like vehicle trajectories, capturing the time-dependent nature of driving behavior and predicting future movements (Wang et al., 2022). CNN-based techniques, on the other hand, extract traffic features from sensor data like images or perceived environment matrices, capturing information about surrounding vehicles, lane markings, and other relevant elements (Feng et al., 2023). Examples include the work of Zhang et al. (2023b), who used a CNN for discretionary lane changes on the highD dataset, and Cheng et al. (2024), who developed a dynamic motion image representation for CNN-based lane change strategy extraction within the SUMO simulator. Recently, attention-based Transformer models have been introduced to further enhance the modeling of temporal and spatial interactions in mixed traffic scenarios (Guo et al., 2024; Li et al., 2025c). For instance, Gao et al. (2023) proposed a dual attention architecture using the Transformer structure that jointly predicts lane change decisions and vehicle trajectories by leveraging attention mechanisms to capture complex dependencies among vehicles. Compared with CNNs and LSTMs, Transformer-based approaches achieve superior accuracy and robustness, particularly in heterogeneous traffic environments.

In the realm of non-deep learning methods, Chen et al. (2024) proposed a macro-micro combined approach that reconstructs multi-lane vehicle trajectories using macroscopic velocity contour maps and microscopic car-following models, effectively integrating characteristics at both macro and micro levels. Sun et al. (2024) introduced the multi-vehicle cooperative control scheme, which mitigates traffic oscillations caused by lane changes through a hierarchical control structure, optimizing overall traffic flow but showing limitations in dynamically adjusting individual vehicles. In contrast, Ma et al. (2025) developed a bidimensional motion planning framework based on the extended omnidirectional risk indicator. By dynamically quantifying real-time risks, this framework adjusts longitudinal and lateral behaviors before, during, and after lane changes, demonstrating superior adaptability and logical consistency, especially in complex, multi-participant traffic environments.

Among existing research, deep learning methods have demonstrated significant advantages in lane-change modeling, primarily due to their powerful feature extraction capabilities. These models effectively capture high-dimensional driving behaviors and model the complex interactions between vehicles and their surrounding environments, excelling in mixed traffic scenarios. However, the inherent lack of interpretability in deep learning models remains a significant challenge, especially for safety-critical applications in intelligent transportation systems. To rigorously evaluate both the performance and interpretability of deep learning models, we select three representative architectures: an LSTM-based model (Wang et al., 2022), a CNN-based model (Zhang et al., 2023b), and a Transformer-based model (Gao et al., 2023). These models are chosen not only for their strong performance, but also because they represent classical and widely adopted neural network architectures, enabling broader generalization of our findings. While most prior studies on deep lane-change models have prioritized prediction accuracy, they have paid limited attention to interpretability under realistic conditions. By introducing our explanation method and evaluating it across LSTM, CNN, and Transformer architectures, we address this gap and provide a practical tool for interpreting model behavior and feature contributions in complex driving scenarios.

3. Attribution computation

3.1. Original Aumann-Shapley attribution

A rigorous understanding of decision-making within DNN-based lane-change models can be achieved through the utilization of Aumann-Shapley (AS) values. These values, rooted in a strong theoretical foundation, adhere to several key interpretability axioms, making them a compelling baseline for attribution computation (Li et al., 2023a; Zhang et al., 2023a; Deng et al., 2024). Specifically, within the context of deep lane-change models, the AS methodology offers a principled approach to quantifying the marginal contribution of each input feature to the final prediction. Our work extends the AS framework, incorporating modifications tailored for lane-change models.

We begin by introducing the definition of AS values within the specific domain of lane-change models. Let $f_d(x)$ represents the output decision of a deep lane-change model, where x denotes the input vector and d signifies a specific maneuver (*e.g.*, left lane change, right lane change, or lane hold). For simplicity, we will omit the subscript d when it does not affect understanding. Given a reference $\bar{x}$, representing the absence of input features (typically selected based on input characteristics to establish relative contributions), the AS value for a given input feature x_i is defined as:

$$\phi_i = \int_{t=0}^{1} \left[f\left(\mu(t) + \frac{\partial \mu_i(t)}{\partial t} \right) - f\left(\mu(t)\right) \right] dt, \tag{1}$$

$$\mu(t) = \bar{x} + t(x - \bar{x}), \tag{2}$$

where $\mu(t)$ represents the integration path for AS value calculation, possessing the same dimensionality as the input x. Here, $\mu(0) = \bar{x}$ denotes the path's origin, and $\mu(1) = x$ denotes its terminus. The fundamental principle underlying AS calculation involves the gradual

restoration of the i-th feature along the path from $\bar{x}$ to x, evaluating the marginal contribution, denoted by ϕ_i, introduced by the absence or presence of the i-th feature along this path. To enhance computational efficiency, the integral is often approximated via a first-order Taylor expansion:

$$f\left(\mu(t)+\frac{\partial \mu_i(t)}{\partial t}\right)=f\left(\mu(t)\right)+\frac{\partial f(\mu(t))}{\partial \mu_i(t)}\frac{\partial \mu_i(t)}{\partial t}+O(dt^2), \tag{3}$$

where the remainder term $O(\cdot)$ encapsulates higher-order corrections, which become negligible as dt approaches zero. Consequently, Eq. (1) can be simplified to the following path integral:

$$\phi_i=\int_{t=0}^{1}\frac{\partial f\left(\mu(t)\right)}{\partial \mu_i(t)}\frac{\partial \mu_i(t)}{\partial t}dt, \tag{4}$$

where the first term of the integrand represents the gradient of the prediction with respect to the i-th feature of the path μ, and the second term denotes the directional derivative of the i-th feature along the path. The resulting attributions ϕ_i represents the marginal contribution of the i-th feature to the prediction.

3.2. Path aggregation attribution

The reliability and accuracy of AS computations are intrinsically linked to the efficacy of the attribution path in traversing the multi-dimensional feature space. Existing AS implementations, predominantly designed for vision models with pixel inputs ranging from 0 to 255, typically utilize one straight-line path characterized by uniformly increasing increments. For example, applying Riemann discretization to Eq. (4) with 101 sample points yields t values of $[0, 0.01, \ldots, 1]$, representing the straight-line path. However, lane-change models present a distinct challenge. Their input features exhibit significantly heterogeneous distributions and possess inherent semantic meaning (*e.g.*, speed difference and headway with the preceding vehicle). This inherent complexity renders the input feature space more intricate than that of image pixels.

Accurate attribution requires sampled points along the attribution path to effectively capture the underlying input feature distribution, thus minimizing noisy attributions. Because identifying a single, universally optimal path is challenging, we explore probabilistic paths to better cover the input feature distribution. Aggregating the attributions derived from various paths allows us to more accurately capture true feature contributions. A brief pipeline of our attribution method is shown in Fig. 2.

Specifically, we sample the intervals β, between path sample points from a Dirichlet distribution $\mathcal{D}$, a member of the exponential family. A key property of the Dirichlet distribution guarantees that these intervals sum to unity. Given K sample points, we define

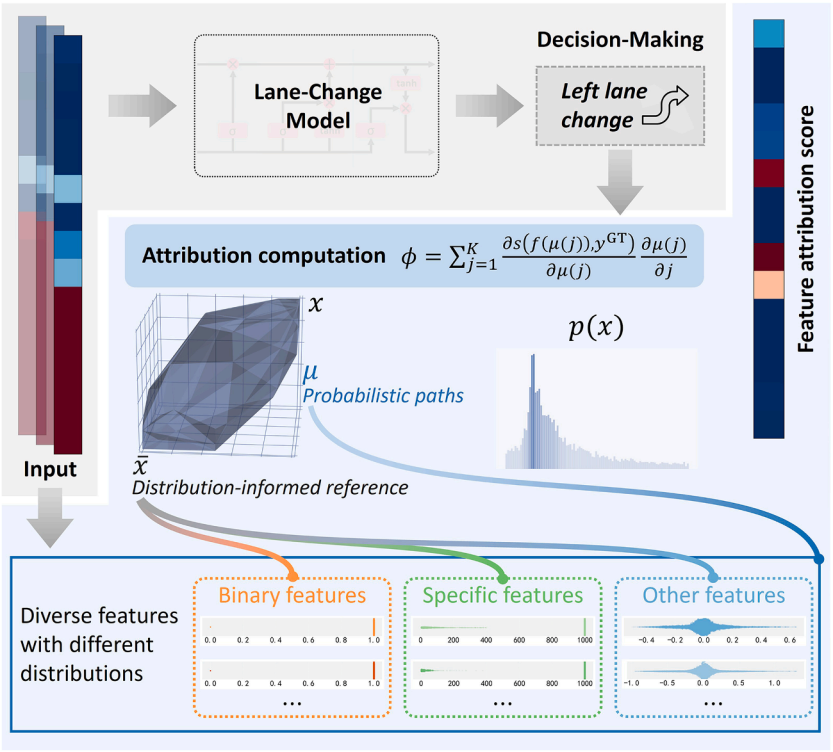

Fig. 2. Pipeline of the proposed attribution method.

the k-th the interpolation point as:

$$B_i(k) = \sum_{j=1}^{k} \beta_j(i), \tag{5}$$

where $B_i(k)$ determines the interpolation point along the path for the i-th feature at the k-th sample point, analogous to the role of t in Eq. (4). Specifically, $B_i(0) = 0$ corresponds to the starting point $\bar{x}_i$, and $B_i(K) = 1$ corresponds to the actual input value x_i. $\beta_j(i)$ represents the path point interval sampled from a Dirichlet distribution $\mathcal{D}(\alpha(i))$, where $\sum_{j=1}^{K} \beta_j(i) = 1$. Although other distributions within the exponential family could offer similar functionality, the Dirichlet distribution is selected for its demonstrated efficacy in generating accurate attributions and the ease of tuning its concentration parameter α in our experiments.

As described by the probability density function of the Dirichlet distribution:

$$f(\beta;\alpha) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{j=1}^{K} \beta_j^{\alpha-1}, \tag{6}$$

where Γ denotes the Gamma function. The concentration parameter α directly controls the distribution of the sampled intervals β. When α is large, the sampled intervals β approach a uniform distribution, resulting in path points that are evenly spaced. Conversely, when α is small, the sampled intervals become uneven, and the probability density increases near the boundaries, where some β_j are close to 0 or 1. Therefore, varying α adjusts the density of the path intervals, enabling the sampling process to better match the actual feature distribution.

To achieve this alignment, we introduce Shannon entropy to quantify the sparsity of each feature's distribution and adaptively assign α for each feature. Features characterized by wider, more sparse distributions are assigned smaller α values, and vice-versa. For sparser distributions, uneven sampling intervals along the path are desirable. Here, a sparse distribution refers to a feature whose values span a wide range, resulting in a relatively uniform histogram. For the i-th feature, we use M histograms to represent the input feature distribution within the dataset as probabilities across different value ranges. Denoting the probability associated with each histogram bin as $p_m(x_i)$, the Shannon entropy for feature i is defined as:

$$H(i) = -\sum_{m=1}^{M} p_m(x_i) \log\left(p_m(x_i)\right), \tag{7}$$

where M is the number of histogram bins. The concentration parameter $\alpha(i)$ for the i-th feature is then defined as:

$$\alpha(i) = \frac{1}{1 + H(i)}. \tag{8}$$

With these definitions, we can now define the Dirichlet-sampled attribution path. For compatibility with neural network algorithms, we discretize the path μ. The value of the i-th feature at the k-th sample point is given by:

$$\mu_i(k) = \bar{x}_i + B_i(k)(x_i - \bar{x}_i). \tag{9}$$

Despite this refined path definition, directly applying AS to deep lane-change models remains problematic due to their vector-valued outputs, which preclude standard backpropagation. To address this issue, we calculate the distance between the model output y and the ground truth label y^{GT} using a Gaussian kernel:

$$s\left(f(x), y^{\mathrm{GT}}\right) = \exp\left(-c\left\|f(x) - y^{\mathrm{GT}}\right\|_2^2\right), \tag{10}$$

where y^{GT} stands for the ground truth label. The parameter c is empirically set to 0.5; its specific value typically does not significantly influence the results. Backpropagation is performed from this distance metric to obtain the gradient information necessary for attribution calculation:

$$\phi_i = \sum_{j=1}^{K} \frac{\partial s\left(f\left(\mu_i(j)\right), y^{\mathrm{GT}}\right)}{\partial \mu_i(j)} \frac{\partial B_i(j)}{\partial j}\left(x_i - \bar{x}_i\right), \tag{11}$$

where j indexes the path sample points and K is the total number of samples in the path. Each sampled set of intervals β, yields a corresponding attribution result ϕ. By sampling multiple sets of intervals, we can generate probabilistic paths within the feature space. The attribution pipeline is illustrated in Fig. 2. The 3D path diagram in Fig. 2 illustrates a convex hull formed by 30 probabilistic paths within a three-dimensional feature space; however, actual calculations occur in a high-dimensional feature space.

The final aggregated attribution is computed as the expectation of the attributions across various paths, estimated by averaging the attribution scores over P paths, where P is the number of generated paths. This aggregation method enhances the accuracy of the feature importance by mitigating the potential impact of any single, potentially suboptimal, path. The procedure provides all the necessary components for attribution calculation, except for the reference $\bar{x}$, the generation method for which is described below.

3.3. Distribution-informed reference

In vision models, where attribution methods are commonly employed, the reference $\bar{x}$ is often set to zero or random noise, representing the absence of meaningful feature information. However, these approaches are ill-suited for deep lane-change models.

Setting features like the relative distance to the preceding vehicle to zero or a random value does not realistically represent the absence of that feature in a driving scenario. Moreover, explicitly defining feature absence within the context of lane-change prediction is inherently challenging. For instance, setting $\bar{x}$ to a specific predefined value, as in previous approaches, is not applicable when inputs comprise ego-vehicle speed and position, along with surrounding traffic information. Defining "absence" for such features becomes conceptually complex.

Given the difficulty of directly defining feature absence, we adopt an implicit definition. We identify $\bar{x}$ by searching for an input that induces a change in the original model decision, thereby indirectly representing feature absence. Drawing inspiration from counterfactual explanations, we optimize $\bar{x}$ to minimize the original lane-change decision score:

$$\mathcal{L}_{\text{score}} = \left| f(\bar{x}) \right|. \tag{12}$$

Since a lane-change model is trained to output a specific lane-change decision, altering the current decision necessarily implies a shift towards another decision. This ensures that the features most relevant to the original decision are effectively absent or minimized in $\bar{x}$.

However, this constraint alone often leads to the generation of adversarial samples. While these samples can effectively alter the model's decision, the resulting $\bar{x}$ often lies outside the distribution of valid inputs. To mitigate this, we introduce a distance loss to constrain the optimization of $\bar{x}$.

For most input features, we apply a constraint based on a threshold τ. Since lane-change models often incorporate binary input features, we employ a near-inversion loss constraint for these features. Furthermore, features like time-to-collision (TTC) are typically encoded as a large value, denoted as $\text{val}(\text{inf})$, when the surrounding vehicle is absent. We introduce an additional loss constraint to handle such specific cases. The final distance loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{dist}} = {} & \mu_1 \sum_{i \in \mathcal{I}_{\text{other}}} \left\| \left| \bar{x}_i - x_i \right| - \tau \right\|_1 + \\ & \mu_2 \sum_{i \in \mathcal{I}_{\text{binary}}} \left\| \bar{x}_i - (1 - x_i) \right\|_1 + \\ & \mu_3 \sum_{i \in \mathcal{I}_{\text{specific}}} \left\| \bar{x}_i - \left(\text{val}(\text{inf}) - \tau \right) \right\|_1, \end{aligned} \tag{13}$$

where μ_1, μ_2, and μ_3 balance the contributions of the different loss terms. $\mathcal{I}_{\text{binary}}$ denotes the set of indices for binary features (*e.g.*, vehicle presence, lane existence). $\mathcal{I}_{\text{specific}}$ denotes the set of indices for features representing infinity (*e.g.*, TTC when no preceding vehicle is present). $\mathcal{I}_{\text{other}}$ represents the indices of all other features. τ is defined as the 10-th percentile of the feature values across all samples in the dataset, calculated independently for each feature dimension. $\text{val}(\text{inf})$ is a large value representing infinity, defined as 999, consistent with the discussion (Zhang et al., 2023b; Wang et al., 2022; Gao et al., 2023). The final reference point $\bar{x}$ is obtained by jointly optimizing $\mathcal{L}_{\text{score}}$ and $\mathcal{L}_{\text{dist}}$.

4. Experiments

4.1. Implementation details

Models and dataset. Our experiments reproduce three distinct deep learning architectures: LSTM-based (Wang et al., 2022), CNN-based (Zhang et al., 2023b), and Transformer-based models (Gao et al., 2023). Both LSTM and CNN models are trained and evaluated on the highD dataset (Krajewski et al., 2018). This dataset comprises trajectories of over 110,500 vehicles, encompassing a diverse range of driving maneuvers, including left and right lane changes and lane-keeping behavior, thus providing a comprehensive platform for assessing our lane-change attributions. The Transformer model is trained on the NGSIM (Afederal Highway Administration, 2021) I-80 and US-101 freeway dataset. The NGSIM dataset is a large-scale, publicly available naturalistic vehicle trajectory dataset collected by the U.S. Federal Highway Administration. It contains high-resolution vehicle trajectory data from two major freeway segments: the I-80 in Emeryville, California, and the US-101 in Los Angeles, California. The dataset covers a diverse range of real-world traffic conditions, including congested and free-flow states, lane-changing maneuvers, and complex vehicle interactions. It provides rich contextual information, such as vehicle types and lane configurations, making it widely used for benchmarking trajectory prediction, discretionary lane-change decision-making, and other intelligent transportation research tasks.

As defined in Zhang et al. (2023b), these models can be classified as discretionary lane-change models. The primary objective of such models is to predict lane-change decision-making at the current moment based on historical trajectories and information about the surrounding environment.

For consistent attribution analysis across lane-change models, the original LSTM model (Wang et al., 2022) is modified by excluding the model predictive control (MPC) component, as our work focuses solely on DNN attribution. Only the neural network part of the model is retained, with the output layer adapted to predict lane-change decisions directly instead of the original MPC-specific outputs. The resulting model exhibits high accuracy, demonstrating the robustness and precision of the LSTM-based architecture.

The LSTM model takes as input a sequence of perceptual information spanning a two-second window preceding the lane change maneuver. This input consists of 15 features: left lane existence (LLaneExist), right lane existence (RLaneExist), lateral relative position difference (RelDiffY), lateral velocity (LatVelY), lateral acceleration (LatAccY), longitudinal velocity (LongVelX), longitudinal acceleration (LongAccX), time-to-collision (TTC) with the preceding vehicle (P-TTC), with the following vehicle (F-TTC), with the left

preceding vehicle (LP-TTC), with the left vehicle (L-TTC), with the left following vehicle (LF-TTC), with the right preceding vehicle (RP-TTC), with the right vehicle (R-TTC), and with the right following vehicle (RF-TTC). These features are provided as sequential data with 50 frames, resulting in a 15×50 input matrix.

The CNN model utilizes the driving operational picture (DOP) as input, a representation of driving style, and outputs three lane-change predictions. The DOP is structured as a 7×8 matrix of seven statistical features calculated across eight vehicle features. The statistical measures, comprising the rows of the input matrix, are: mean, standard deviation, median, 25-th percentile, 75-th percentile, minimum, and maximum. The columns represent the following vehicle features: relative longitudinal location (RelLocX), relative lateral location (RelLocY), longitudinal velocity (LongVelX), lateral velocity (LatVelX), longitudinal acceleration (LongAccY), lateral acceleration (LatAccY), space headway distance (HeadDist), and time headway (HeadTime). The vehicles considered for feature extraction include the ego vehicle (Ego), preceding vehicle (P), preceding vehicle in the left and right adjacent lanes (LP and RP), following vehicle in the left and right adjacent lanes (LF and RF), and alongside vehicle in the left and right adjacent lanes (L and R). Thus, the input to the lane-change decision-making model consists of eight DOP matrices. Following the approach described in the original paper, we also employ 8×8×7 = 448 features. Both the LSTM and CNN models incorporate a two-second reaction time assumption for lane-change decisions, reflecting typical human driver behavior. Specifically, the input sequences span a two-second window preceding the lane change maneuver, as described in Zhang et al. (2023b). Furthermore, Ali et al. (2023) have advocated for the continuous usage of data. Integrating methods for continuous data processing may further improve the performance of these lane-changing models.

In addition to the above models, we also implement the lane-change decision-making component of the Transformer-based architecture proposed by Gao et al. (2023). This model is specifically designed to predict the probability distribution over lane-change decisions by leveraging the attention mechanism to capture both the temporal evolution of the ego vehicle's behavior and the spatial interactions with surrounding vehicles. For each sample, the input comprises the historical positions of the ego vehicle, as well as interaction features from up to five surrounding vehicles over an 18 s observation window (with virtual vehicles introduced when fewer than five are present). The selection of an 18 s time window follows the findings of Gao et al. (2023), which demonstrated that this configuration achieves optimal predictive performance. The interaction features include the longitudinal distance and velocity difference between the ego vehicle and the preceding vehicle (P-Dist, P-Vel), left preceding vehicle (LP-Dist, LP-Vel), left following vehicle (LF-Dist, LF-Vel), right preceding vehicle (RP-Dist, RP-Vel), and right following vehicle (RF-Dist, RF-Vel). These features collectively capture the spatial-temporal relationships and dynamic interactions essential for decision-making. All input features are normalized according to the preprocessing procedures described in the original study to ensure consistent scaling and mitigate the influence of varying units. The decision prediction module employs a multi-head self-attention mechanism to effectively extract correlations among the ego and surrounding vehicles, and outputs a lane-change decision probability vector via a fully connected layer with softmax activation.

Our research focuses on interpreting and analyzing the decision-making mechanisms of these existing lane-change models. To ensure consistency with the original studies, we adopt the data processing, model architectures, and model training workflows described in Wang et al. (2022), Zhang et al. (2023b) and Gao et al. (2023), as well as the corresponding open-access codes, without making any optimizations or modifications to the neural network structures, feature selections, or data processing procedures. This approach helps guarantee the reproducibility of our experiments and the fairness of the comparative analysis.

Parameter selection. A key parameter in our method is the number of paths P. Setting $P = 1$ approximates the original Aumann-Shapley (AS) method, utilizing a single probabilistic path for attribution calculation. For $P > 1$, the attribution computation employs P paths, aggregating the resulting attributions. Empirically, $P = 30$ can represent a point of diminishing returns for attribution accuracy across both evaluated lane-change models. Although this value proved effective in our experiments, manual adjustment may be necessary when applying to other models.

The concentration parameter α in Eq. (8) can be scaled to control dispersion and prevent extreme values. For instance, scaling by a factor b yields:

$$\alpha(i) = \frac{1 + bH(i)}{1 + H(i)}, \tag{14}$$

where the scalar b requires adjustment based on the specific path under consideration, with zero often serving as a suitable default value. However, in cases where the path-generated attribution exhibits instability, a small increase in b (*e.g.*, $b = 0.1$), can enhance robustness.

Regarding the loss balancing during the reference $\bar{x}$ optimization, we find that stable optimization can be achieved when the three terms in Eq. (13) exhibit comparable magnitudes, while being smaller than Eq. (12). Therefore, we set $\mu_1 = 0.04$, $\mu_2 = 0.2$, and $\mu_3 = 0.02$, and use the Adamax optimizer. Early stopping is implemented when the loss variation remains below 5 % within 10 epochs.

Other parameters are less sensitive and generally require no further tuning. The number of discretization points K is set to 50. Values of K significantly smaller than 50 result in reduced attribution accuracy, while larger values offer negligible improvement. The number of histogram bins M is set to 100, with no apparent benefit observed from using larger values.

4.2. Quantitative comparison of attribution methods

In this section, we present a quantitative comparison between our attribution method and several state-of-the-art methods, applied to three lane-change models (Wang et al., 2022; Zhang et al., 2023b; Gao et al., 2023). We evaluate performance using four complementary metrics: sensitivity-n (Sen-n) (Schulz et al., 2020), accuracy information curve (AIC), softmax information curve (SIC)

Table 1
Quantitative attribution comparison on the LSTM model.

	Sen-n ↑	AIC ↓	SIC ↓	MRF ↓
PRCA	0.667	0.283	0.264	0.169
IDGI	0.621	0.241	0.239	0.142
WB-LRP	0.662	0.294	0.285	0.177
SVCE	0.717	0.217	0.209	0.109
GTCAM	0.702	0.234	0.207	0.119
SAMP	0.669	0.238	0.211	0.127
$10\mu_1$	0.729	0.188	0.197	0.099
$10\mu_2$	0.712	0.219	0.208	0.114
$10\mu_3$	0.709	0.212	0.211	0.112
$0.1\mu_1$	0.717	0.191	0.202	0.104
$0.1\mu_2$	0.714	0.202	0.203	0.111
$0.1\mu_3$	0.735	0.185	0.195	0.091
ZeroRef	0.716	0.206	0.191	0.114
LinePath	0.684	0.267	0.252	0.137
$P = 1$	0.682	0.269	0.258	0.144
$P = 10$	0.722	0.181	0.194	0.103
$P = 45$	**0.769**	**0.171**	0.189	0.083
$P = 50$	0.759	0.175	0.188	0.086
$P = 30$ (Ours)	0.765	0.172	**0.186**	**0.082**

(Kapishnikov et al., 2019), and most-relevant-first (MRF) (Schulz et al., 2020). Sen-n measures the efficiency axiom, a fundamental aspect of theoretical soundness, requiring that attribution sums equal the decision score. The remaining metrics (AIC, SIC, and MRF) quantify attribution accuracy by measuring changes in decision scores after removing important features. While AIC and SIC leverage information entropy for this assessment, MRF directly utilizes the decision scores themselves. Together, these metrics provide a comprehensive evaluation framework.

We compare our method against several contemporary attribution methods: PRCA (Chormai et al., 2024), IDGI (Yang et al., 2023), WB-LRP (Li et al., 2025d), SVCE (Li et al., 2024b), GTCAM (Li et al., 2024d), and SAMP (Zhang et al., 2024). PRCA and WB-LRP employ layer-wise relevance propagation techniques. For LSTM application, we treat the LSTM module as an integrated combination of linear and non-linear layers, distributing relevance proportionally to weights as the propagation rule. The CNN model, predominantly using linear and ReLU activations, requires no specific rule adjustments. When applied to the Transformer architecture, LRP propagates the relevance scores backward through the multi-head self-attention layers, fully connected layers, and any other network components, following the structure of the model. To propagate relevance through the attention layer, the relevance assigned to each output token is redistributed to the input tokens based on the computed attention weights. The model-agnostic methods (IDGI, SVCE, and SAMP) require minimal modification for implementation. We adapt SVCE, originally designed with Shapley values for reinforcement learning lane-change models, by implementing random sampling for computational efficiency. GTCAM, a Shapley-value-based method for graph neural networks, requires a predefined partition matrix for CNN application. We partition the input based on vehicle locations, *i.e.*, ego, preceding, following, left, left preceding, left following, right, right preceding, and right following. This approach simplifies the original method, which utilized anatomical partitions designed for a skeleton-based behavior recognition model, while remaining practical for traffic scenarios that naturally exhibit inherent spatial relationships.

Tables 1–3 present the quantitative results. Our method consistently outperforms competitors on the LSTM model. Among Shapley-value-based methods, SVCE achieves the second-best performance, underscoring the effectiveness of Shapley values for decision attribution. PRCA and WB-LRP show relatively poor performance, likely because their original formulations are not specifically designed for LSTMs, suggesting the need for theoretical refinements when applying these methods to new architectures. On the CNN model, our method also maintains superior performance. Among these tested methods, GTCAM demonstrates strong results. This may be attributed to the model input's natural partitioning of surrounding vehicle information, which aligns well with GTCAM's fundamental design principles. It is important to note that our proposed method does not employ any explicit partitioning strategy; instead, it processes the entire data holistically, without predefined groups or consortia. In contrast, among the compared methods, only GTCAM relies on partitioning strategies based on semantic priors, where the input is divided according to domain knowledge. Such partitioning allows GTCAM to capture more flexible and context-aware interactions between different regions or entities, rather than being limited by rigid, manual partitions. Our experimental results indicate that prior-based partitioning can improve performance for LSTM and CNN architectures, likely due to the structural alignment between the partitioning strategy and these models. However, this improvement was not observed with the Transformer model, suggesting that its performance may be sensitive to model architectures.

On the Transformer model, our method achieves the best overall attribution performance, as shown in Table 3. These results demonstrate that our approach is well-suited for interpreting decision processes in attention-based architectures. Notably, Shapley-value-based methods such as SVCE and SAMP also perform strongly, indicating their robustness in high-capacity models with complex input dependencies. In contrast, relevance propagation methods like PRCA and WB-LRP yield lower AIC, SIC, and MRF scores, which may be due to the unique information flow and multi-head aggregation mechanisms inherent to Transformers. Overall, the

Table 2
Quantitative attribution comparison on the CNN model.

	Sen-n ↑	AIC ↓	SIC ↓	MRF ↓
PRCA	0.624	0.341	0.359	0.178
IDGI	0.562	0.366	0.352	0.187
WB-LRP	0.551	0.394	0.405	0.199
SVCE	0.512	0.377	0.392	0.189
GTCAM	0.619	0.278	0.289	0.146
SAMP	0.652	0.282	0.291	0.152
$10\mu_1$	0.675	0.263	0.271	0.139
$10\mu_2$	0.603	0.285	0.297	0.149
$10\mu_3$	0.625	0.275	0.284	0.147
$0.1\mu_1$	0.658	0.269	0.283	0.142
$0.1\mu_2$	0.629	0.271	0.266	0.143
$0.1\mu_3$	0.693	0.254	0.263	0.131
ZeroRef	0.633	0.267	0.253	0.149
LinePath	0.601	0.299	0.282	0.171
$P = 1$	0.598	0.308	0.312	0.183
$P = 10$	0.682	0.262	0.274	0.143
$P = 45$	**0.707**	0.249	0.255	0.124
$P = 50$	0.697	0.247	0.257	**0.122**
$P = 30$ (Ours)	0.706	**0.242**	**0.251**	0.128

Table 3
Quantitative attribution comparison on the Transformer model.

	Sen-n ↑	AIC ↓	SIC ↓	MRF ↓
PRCA	0.572	0.308	0.317	0.176
IDGI	0.628	0.283	0.291	0.153
WB-LRP	0.563	0.306	0.319	0.184
SVCE	0.643	0.269	0.278	0.142
GTCAM	0.609	0.286	0.295	0.162
SAMP	0.662	0.255	0.267	0.131
$10\mu_1$	0.713	0.224	0.238	0.109
$10\mu_2$	0.668	0.248	0.256	0.129
$10\mu_3$	0.681	0.237	0.249	0.124
$0.1\mu_1$	0.702	0.233	0.244	0.115
$0.1\mu_2$	0.688	0.229	0.243	0.119
$0.1\mu_3$	0.741	0.211	0.218	0.094
ZeroRef	0.693	0.225	0.238	0.116
LinePath	0.639	0.272	0.283	0.146
$P = 1$	0.621	0.287	0.297	0.158
$P = 10$	0.696	0.239	0.251	0.118
$P = 45$	0.749	0.203	0.209	0.092
$P = 50$	0.751	0.197	0.213	**0.087**
$P = 30$ (Ours)	**0.756**	**0.195**	**0.203**	0.089

quantitative results suggest that our method provides more faithful and stable attributions on the Transformer model, further validating its generalizability across diverse neural network architectures.

4.3. Ablation studies

In this section, we conduct a series of ablation studies to evaluate our method. We first investigate the reference generation process, focusing on the impact of three key hyperparameters: μ_1, μ_2, and μ_3 (set to 0.04, 0.2, and 0.02, respectively, in our original setting). For each parameter, we assess the effects of increasing its value by a factor of 10 and decreasing it to 0.1 times its original value, while keeping the remaining parameters fixed. The results for the LSTM, CNN, and Transformer models are presented in Tables 1–3, respectively.

For the LSTM model, increasing μ_1 to $10\mu_1$ yields a Sen-n of 0.729 and an AIC of 0.188, whereas decreasing it to $0.1\mu_1$ still maintains strong performance (Sen-n = 0.717, AIC = 0.191), both outperforming most baselines. Similar trends are observed for μ_2 and μ_3, with $0.1\mu_3$ achieving the best sensitivity (0.735) and the lowest AIC (0.185), suggesting that smaller values for μ_3 may be particularly advantageous in some cases. On the CNN model, $0.1\mu_3$ also achieves the highest Sen-n (0.693), as well as the lowest AIC (0.254) and MRF (0.131) among all ablation settings, while other hyperparameter variations continue to outperform most competing

methods. For the Transformer model, similar trends are observed as in LSTM and CNN. A smaller value of μ_3 (*e.g.*, $0.1\mu_3$) achieves the best Sen-n and the lowest AIC. Additionally, variations in μ_1 and μ_2 still outperform most baselines, demonstrating the robustness of the proposed method to these hyperparameters.

These findings indicate that our method is relatively robust to the choice of μ_1, μ_2, and μ_3, but changing the numerical ranges of the loss terms leads to performance degradation. In general, it is advisable to select hyperparameters such that the values of these terms are kept within comparable ranges to achieve the best results. Across all three model architectures, our setting yields the best overall performance, underscoring the effectiveness and stability of our hyperparameter selection strategy for reference generation.

We additionally compare different reference points. As most existing methods are primarily designed for vision models, several effective techniques, such as random noise, image alpha blending, mask optimization, and minimalist baselines (Singla et al., 2023; Mamalakis et al., 2023), are not suitable for lane-changing models. Thus, our comparison is limited to the zero reference point (ZeroRef), with inversion operations applied to binary features. Although the choice of reference point has less impact than the attribution path, it still noticeably influences attribution results.

Furthermore, we conduct ablation studies focusing on the attribution path, using the original Aumann-Shapley straight-line path (LinePath) as a baseline. We evaluate different path numbers, $P = (1, 10, 30, 45, 50)$. When $P = 1$, which corresponds to a single random path without aggregation, the results are inferior to LinePath and inadequately capture feature transitions from absence to presence. Performance improves significantly when aggregating multiple probabilistic paths, plateauing at approximately $P = 30$. Increasing P to 40 or 50 yields negligible additional gains, supporting our selection of 30 paths as an optimal balance between performance and computational efficiency.

4.4. Qualitative attribution analysis on LSTM Model

In this section, we use an LSTM model to perform attribution analysis in a lane-change scenario. Fig. 3 illustrates a left lane-change scenario and its corresponding attributions. At the top of the figure, the overall traffic scene is shown, captured by a drone-mounted camera, presenting the entire lane-change maneuver. The ego vehicle (ID14, in red) transitions from the rightmost lane to the middle lane. Below, the attribution results for each method are displayed, corresponding directly to the input features. Dark blue corresponds to small values, while red indicates large values.

The input feature vector depicted in the figure is extracted from the traffic scene information. This vector encompasses 50 frames of data, summed for visualization. The lane-change model bases its decision on these features. Note that the 50 frames of input features do not correspond directly to the scene depicted above but rather to the 50 frames recorded before the lane-change maneuver. In the input feature vector, dark red blocks represent large values. For time-to-collision (TTC) features here, this indicates an infinite time to collision with the corresponding vehicle (represented as 999 in programming practice). From the visualization, we observe that the following vehicle (ID18) is traveling at a higher velocity, potentially leading to a collision if both vehicles maintain their current trajectories. Similarly, the ego vehicle is significantly faster than the preceding vehicle (ID9), also suggesting a possible collision. Displacement and velocity features, being relatively smaller compared to the TTC values, are shown as dark blue in the input feature vector.

Based on this input, the LSTM model outputs a left lane-change decision. We apply various attribution methods to this decision to compute attribution scores, as shown in the lower part of Fig. 3. Most methods highlight the following as key factors:

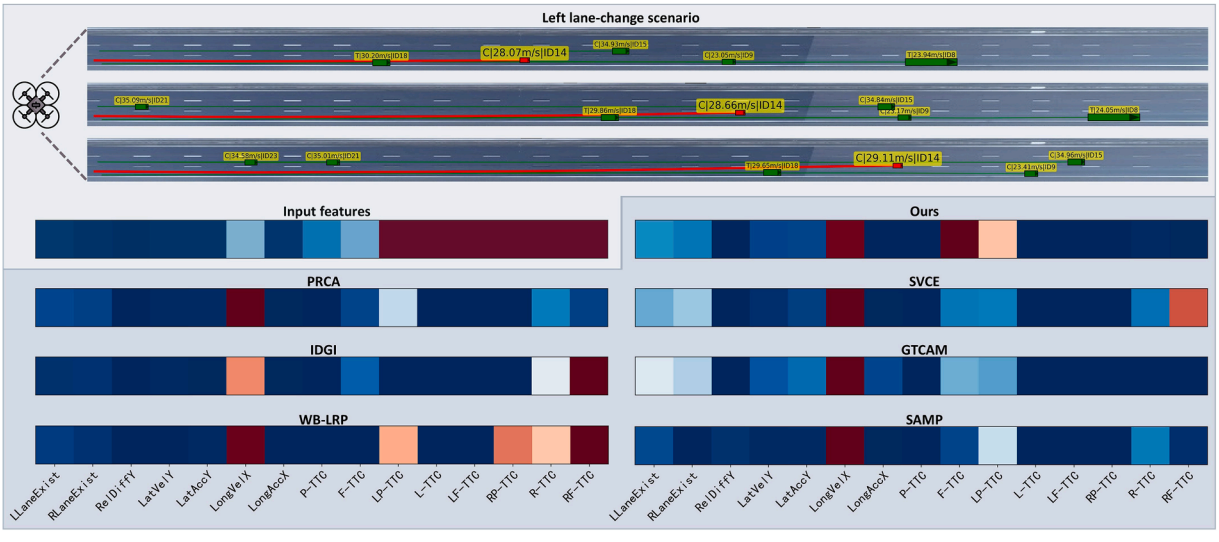

Fig. 3. Attribution explanations generated by seven different methods for the LSTM model.

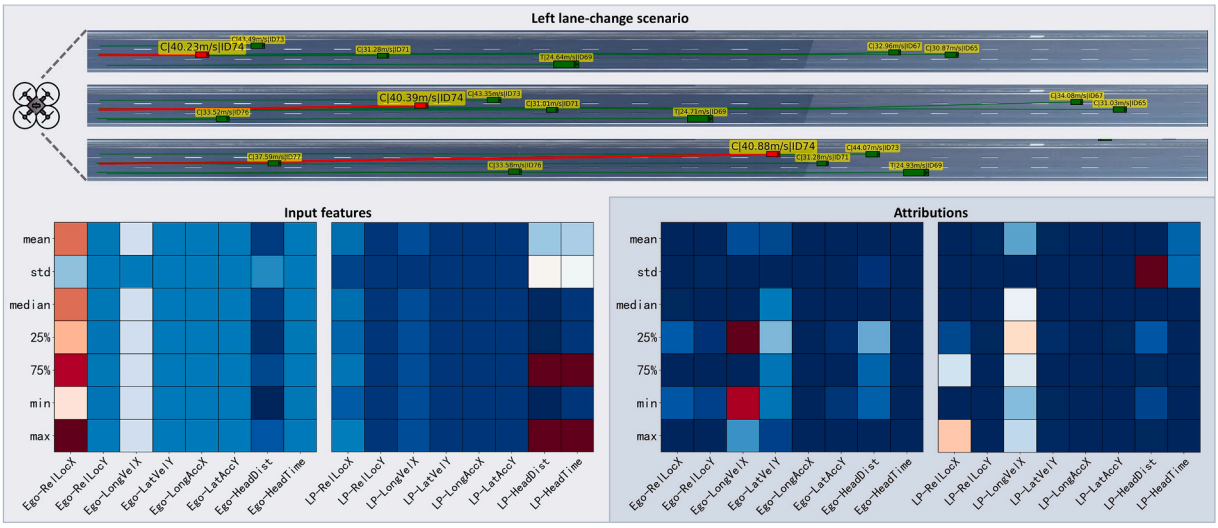

Fig. 4. Attribution explanations for the CNN model.

time-to-collision with the following vehicle (F-TTC), with the left preceding vehicle (LP-TTC), longitudinal velocity (LongVelX), and lane existence (LLaneExist and RLaneExist). Although many methods correctly identify the importance of LongVelX, F-TTC, and LP-TTC, some methods (excluding GTCAM and our method) assign high importance to the time-to-collision with the right following vehicle (RF-TTC) and the right vehicle (R-TTC). This is counterintuitive, as introducing a hypothetical vehicle to the right of the ego vehicle (making R-TTC and RF-TTC finite) would not alter the lane-change decision. This suggests that the high attribution scores for R-TTC and RF-TTC from these methods may be inaccurate. Additionally, the LLaneExist or RLaneExist features clearly contribute less than the TTC features, and our method also effectively captures this distinction.

Our attribution results demonstrate some alignment between the model's feature processing and human intuition. We note that the ego vehicle's speed is higher than the preceding vehicle's and lower than the following vehicle's, creating a risky situation if no lane change is performed. This information is also essential for human drivers. Furthermore, the model considers factors relevant to a left lane change, such as the speed of the left preceding vehicle (allowing sufficient space for the maneuver) and the presence of vehicles to the left following (ensuring a safe following distance). This alignment between human intuition and the attribution results suggests that the lane-change model, trained on real-world driving data, learns certain aspects of human driving behavior. These observations are currently qualitative; a more rigorous quantitative analysis using attributions to assess the alignment between the model and human attention is presented in Section 4.11.

4.5. Attribution explanations on CNN model

In this section, we use a CNN model for attribution analysis. Fig. 4 shows a lane-change scenario where the red car (ID74) represents the ego vehicle. The top part of the figure illustrates the traffic environment during the lane-change maneuver. The bottom left shows a subset of the model's input features contributing to the model's decision, while the bottom right presents the corresponding attribution explanations. Dark blue corresponds to small values, while red indicates large values. Although the model input consists of information from multiple surrounding vehicles, for clarity, we only display the two vehicles that contribute the most to the decision: the Ego and left preceding (LP) vehicle information. Note that the scene depicted in the figure does not correspond directly to the input features but serves as a visual representation of the scenario. The lane-change decision is based on information preceding the lane-change maneuver.

In this traffic scene, the model decides to change to the left lane. Our attribution explanations clearly highlight the reasoning behind this decision. The primary influencing factors are the 25-th and 75-th percentiles of the ego vehicle's longitudinal velocity (Ego-LongVelX), along with its lateral velocity (Ego-LatVelY). These values indicate the ego vehicle's gradual acceleration and its tendency to move leftward. Additionally, the longitudinal velocity of the left preceding vehicle (LP-LongVelX) and its headway distance (LP-HeadDist) from the ego vehicle also contribute to the decision. From a human driver's perspective, these features directly reflect the temporal and spatial margins available for executing a safe left lane change.

The attribution results reveal that the deep model's decision-making process closely aligns with human intuition. Notably, for left lane changes, the contributing features consistently pertain to the ego vehicle and the left preceding vehicle. This observation is consistent across both the LSTM and CNN models. Despite having fundamentally different architectures and training paradigms, these two models exhibit similar feature extraction capabilities. This finding motivates us to conduct further experiments to explore the statistical feature contributions in Section 4.7, and the alignment between model behaviors and human intuition in Section 4.11.

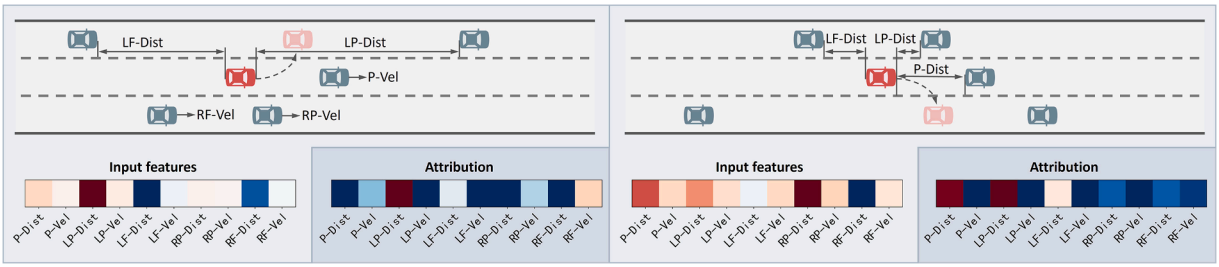

Fig. 5. Attribution explanations for the Transformer model.

4.6. Attribution explanations on transformer model

In this section, we present attribution analysis results based on the Transformer model. Fig. 5 illustrates two representative lane-change scenarios, with the red car denoting the ego vehicle. The upper part of each subfigure visualizes the traffic environment during the maneuver. The visualization format of the traffic scenarios in Fig. 5 is inspired by Gao et al. (2023). The bottom left of each subfigure shows the model's input features, and the corresponding attribution explanations are displayed to the right. Color encoding is consistent with previous figures: dark blue denotes low feature values, while red indicates high feature values.

In the left scenario, the ego vehicle initiates a lane change to the left. The attribution results indicate that the most influential features are the longitudinal velocity differences between the ego vehicle and the preceding vehicle (P-Vel), the right-preceding vehicle (RP-Vel), and the right-following vehicle (RF-Vel), as well as the distances to the left-preceding vehicle (LP-Dist) and the left-following vehicle (LF-Dist). The high attribution scores for LP-Dist and P-Vel suggest that the Transformer model assigns considerable importance to the available spatial gap in the target lane and the ego vehicle's current motion state when making lane-change decisions. Additionally, the influence of RP-Vel and RF-Vel implies that the model takes into account the velocity of vehicles in the adjacent lane, likely to ensure a safe maneuver.

In the right scenario, where a lane change to the right is depicted, the attribution map highlights similar patterns. The primary contributing features are P-Dist, LP-Dist, and LF-Dist. Notably, the Transformer model attributes substantial importance to the spatial distance of neighboring vehicles in the lane on the other side, reflecting a cautious decision-making policy that aligns well with human driving intuition.

Across both scenarios, the attribution results indicate that the Transformer model effectively captures the critical temporal and spatial features necessary for safe lane changes. Compared to the CNN and LSTM models, the Transformer demonstrates a more distributed attention pattern, considering a broader set of surrounding vehicles and their dynamic interactions. This observation suggests that the self-attention mechanism in the Transformer enables the model to integrate information from multiple sources, potentially leading to more robust and context-aware decisions.

4.7. Attribution-based feature contribution analysis

This section presents a statistical analysis of the model input features, leveraging attribution results. Initially, we sum the attribution scores for each input feature across the entire dataset. This summation yields the contribution proportion of each feature, offering a clear perspective on the model's assessment of feature importance. Subsequently, we employ scatter plots to visualize the relationship between feature values and their corresponding attribution scores. This visualization allows us to observe the influence patterns of different features on the decision-making process.

Fig. 6 illustrates the analysis results for the LSTM model. The upper part of the figure reveals that lateral acceleration (LatAccY) is the only feature with a comparatively large contribution; the contributions of other features are relatively uniform. This suggests a dispersed influence of input features in the LSTM model, with no single dominant feature emerging. The lower part of Fig. 6 displays three features exhibiting more prominent influence patterns. For time-to-collision (TTC), significant attributions (both positive and negative) are typically observed when TTC values are either very large (indicating no collision risk) or very small (indicating a potential collision if current states are maintained). Regarding longitudinal velocity (LongVelX), only speeds exceeding a certain threshold have a substantial impact on the decision; lower speeds (from 20 m/s to 30 m/s) exhibit notably smaller attribution values. These findings indicate that the deep model has learned to prioritize some critical feature information.

Fig. 7 presents the corresponding analysis for the CNN model. In contrast to the LSTM model, the CNN model exhibits significant variations in the contributions of different features. The combined contribution of longitudinal velocity (LongVelX) and relative longitudinal location (RelLocX) accounts for more than half of the total attribution. Certain features, primarily relative lateral location (RelLocY) and longitudinal acceleration (LongAccX), make virtually no effective contribution. This disparity likely arises from the different input feature design strategies. The LSTM model uses raw physical quantities directly, whereas the CNN model's input incorporates statistical information, such as mean and standard deviation. RelLocY and LongAccX in the CNN model exhibit minimal variation within the 2s time sequence, leading to insignificant statistical features.

The lower part of Fig. 7 displays scatter plots of input features and their attributions. These plots reveal that for LongVelX, the absolute value of the input feature must exceed a certain threshold to employ a significant influence on the decision. Headway distance

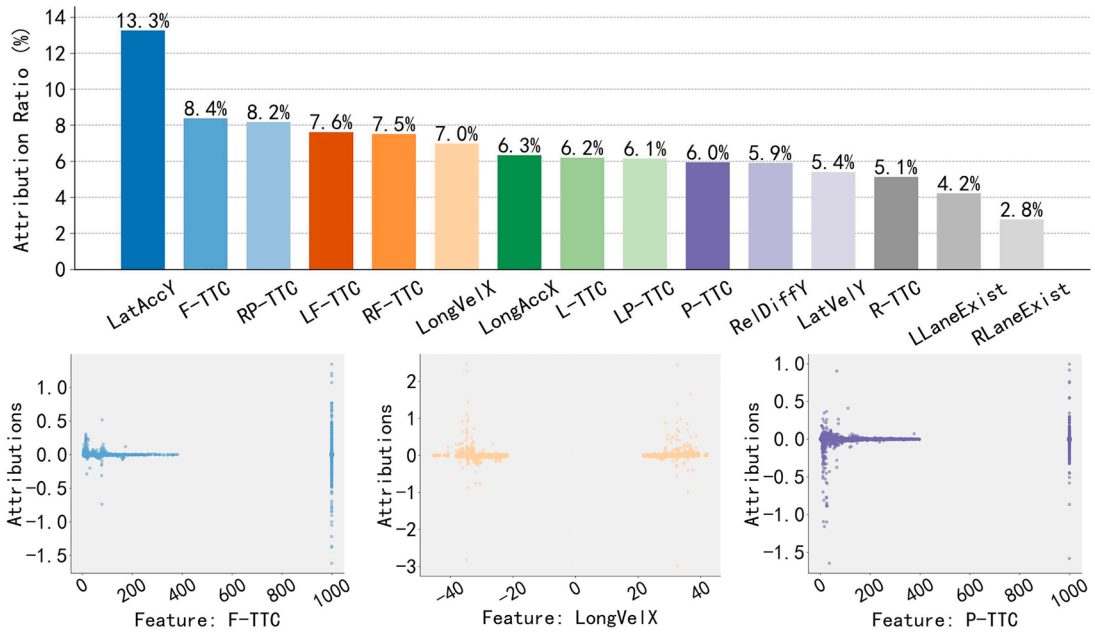

Fig. 6. Contribution analysis of input features in the LSTM model.

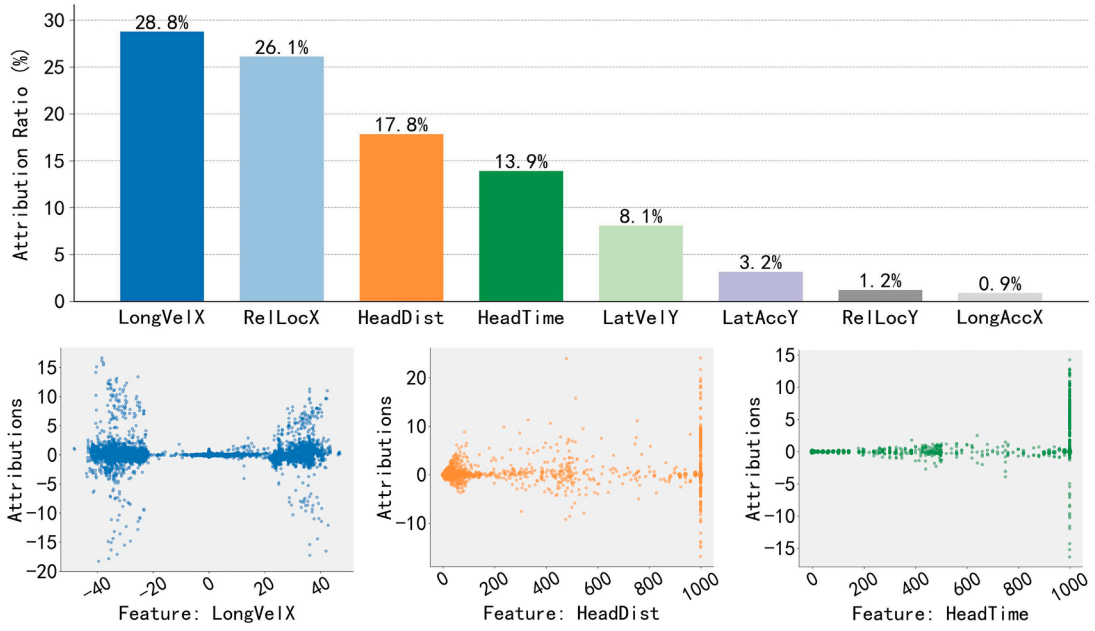

Fig. 7. Contribution analysis of input features in the CNN model.

(HeadDist) and headway time (HeadTime) exhibit patterns analogous to the TTC features in the LSTM model. When headway time is infinite, the contribution tends to be larger than at other values. This situation typically occurs when a relevant vehicle is absent or the relative velocity between vehicles poses no collision risk. Such information is also crucial for human drivers when making lane-change decisions.

Fig. 8 shows the attribution analysis results for the Transformer model. Unlike the CNN model, the Transformer assigns relatively balanced importance across a broad set of features. The top contributing features, such as LF-Dist, RP-Dist, and RF-Dist, each account for around 12% of the total attribution, and even the least important features retain non-negligible influence. This indicates that the Transformer model considers information from multiple surrounding vehicles simultaneously. The scatter plots in the lower part of Fig. 8 further reveal that lower absolute values of distance features are generally associated with higher attribution scores. This

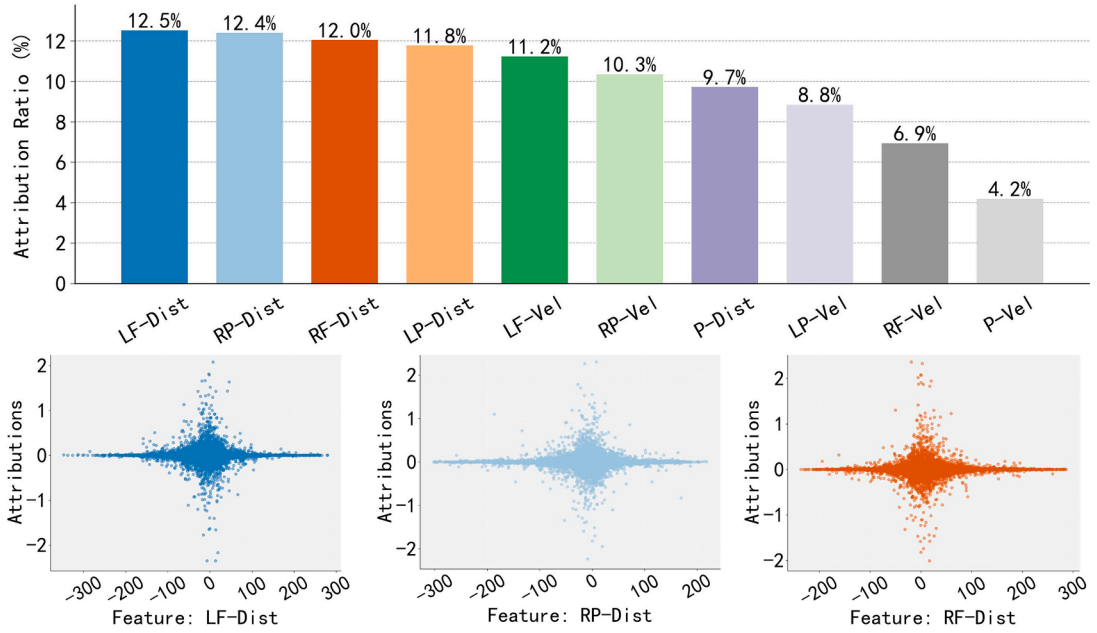

Fig. 8. Contribution analysis of input features in the Transformer model.

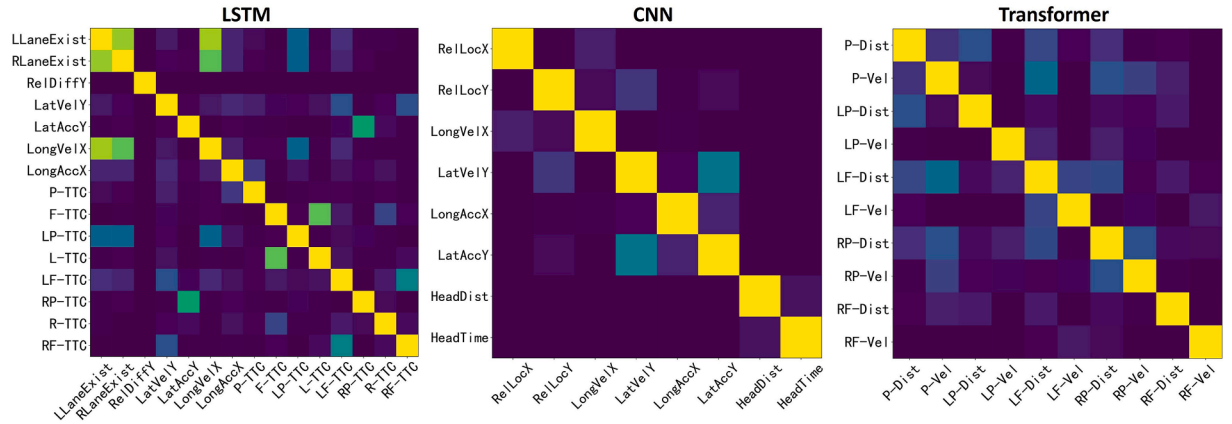

Fig. 9. Similarity of input feature contributions to decision-making in the deep lane-change models measured by centered kernel alignment.

suggests that the Transformer model is sensitive to significant spatial relationships when making lane-change decisions, leveraging its self-attention mechanism to capture complex interactions in the traffic scene.

Drawing inspiration from research on neuron feature similarity in deep models (Shi et al., 2024; Cortes et al., 2012; Kornblith et al., 2019), we further employ attribution values to compute centered kernel alignment (CKA) to measure the similarity of feature contribution patterns. Fig. 9 demonstrates that the contribution patterns of most features are dissimilar, with some exceptions. In the LSTM model, the contribution patterns of the two lane-existence features consistently exhibit high similarity. In other words, when the existence of a left lane significantly contributes to a decision in a given traffic scenario, the existence of a right lane typically has a comparable contribution. Furthermore, we observe a strong similarity between the influence of LongVelX and lane existence. To validate this finding, we remove the two lane existence features from the dataset and retrain the LSTM model. The lane-change prediction accuracy decreased only marginally, from 92.6 % to 92.5 %. This suggests that the lane existence features may be redundant, and removing them could potentially simplify the model architecture. For the CNN model, the feature contributions appear less correlated, suggesting that these features are more independent. For the Transformer model, the CKA analysis shows that most feature contribution patterns are relatively uncorrelated, similar to the CNN model, though with slightly more moderate similarities between certain distance and velocity features. This indicates that the Transformer, leveraging its self-attention mechanism, flexibly integrates information from multiple sources without over-relying on specific feature pairs. Overall, the Transformer achieves a balanced and context-aware attribution pattern, supporting robust decision-making.

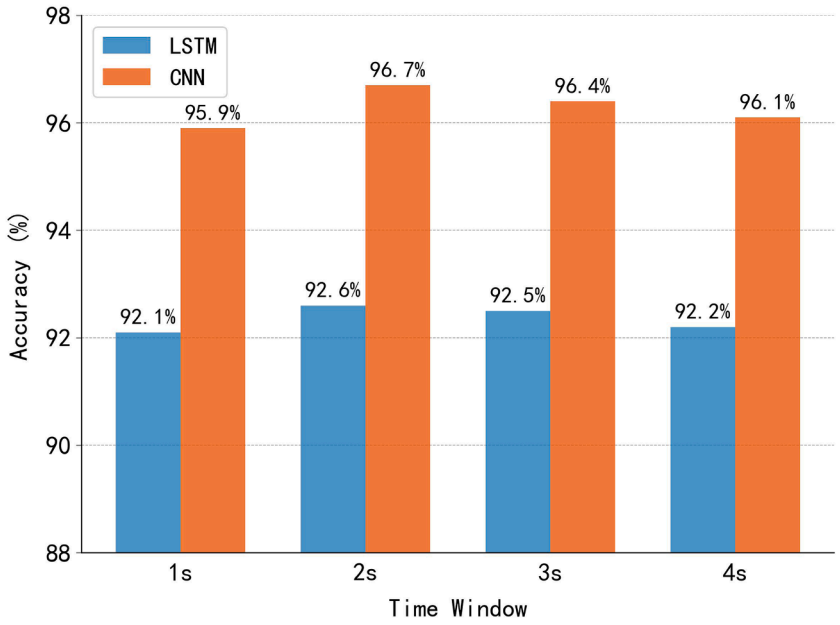

Fig. 10. Model accuracies with different time window sizes.

4.8. Impact of time window size

In this section, we systematically analyze the effects of different time window sizes on model performance and attribution results by designing comparative experiments with varying time window configurations. Specifically, we select 1-, 2-, 3-, and 4 s time windows for independent training and evaluation of both CNN and LSTM models. It is important to note that in the highD dataset, the key information related to lane-changing behavior mainly concentrates within the 3s preceding the maneuver. As the time window increases, the number of valid samples decreases significantly; thus, we do not consider window sizes longer than 4s in this study.

The experimental results are shown in Fig. 10. Both the CNN and LSTM models achieve the highest prediction accuracy when the time window is set to 2 s. Although different time windows only slightly influence the overall model performance, the shortest window (1 s) consistently yields the lowest accuracy for both models. This may result from insufficient information being captured before the lane change, which negatively affects the model's decision process. Overall, these findings confirm the rationality and robustness of adopting a 2 s time window as the default setting.

To further analyze how different time window sizes affect the model's decision mechanism, we conduct frame-wise and feature-wise attribution analysis for the best and second-best models trained with different window sizes. The attribution scenario corresponds to Fig. 3. For the LSTM model, regardless of the time window used during training, the frames immediately preceding the lane change always make the greatest contributions to the model's decision. Fig. 11 illustrates the feature attributions for the three most contributive frames (*i.e.*, the three frames immediately before the lane change). The attribution analysis shows that the LSTM models maintain consistent feature contribution patterns under different window settings, with time-to-collision with the following vehicle (F-TTC) and with the left preceding vehicle (LP-TTC) always being the most influential features for decision-making.

Similarly, as shown in Fig. 12, the CNN model also presents similar attribution results across different time windows, with the traffic scenario identical to that in Fig. 4. The model's attention consistently focuses on the key kinematic and spacing information between the ego vehicle and the left preceding vehicle. For example, F-TTC and LP-TTC remain the most contributive features in the CNN attribution analysis. These results indicate that changes in time window size do not substantially alter these models' decision mechanism. Meanwhile, the attribution analysis further validates the robustness of the proposed interpretability method across different models and time window settings.

4.9. Feature dependency and sensitivity analysis

In this section, we investigate the dependency and contribution scope of different models to feature subsets through a systematic feature ablation study. Inspired by the previous study (Ali et al., 2022), this experiment adopts the recursive feature elimination method (Chen and Jeong, 2007) to systematically analyze the performance of CNN, LSTM, and Transformer models on varying feature subsets.

The experimental results in Fig. 13 indicate that all models experience significant performance limitations when the number of features is small. As the number of features increases, the model performance improves steadily. However, the optimal feature set does not necessarily correspond to the full feature set. For example, the LSTM model originally receives 15 input features. When 5 features

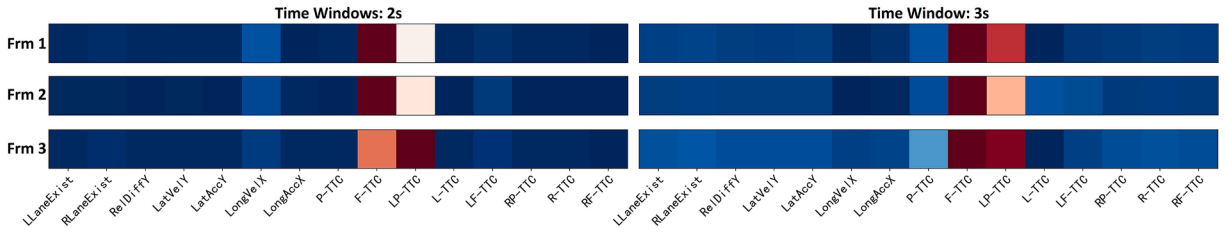

Fig. 11. Attribution results of the LSTM models trained with different time window sizes.

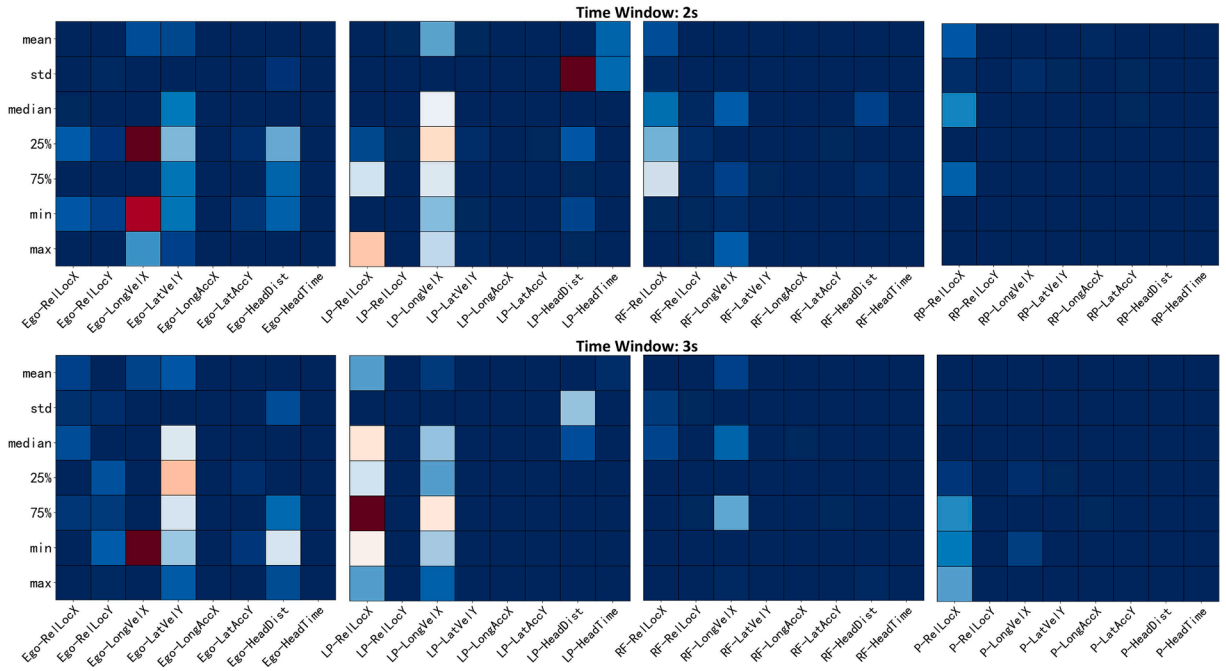

Fig. 12. Attribution results of the CNN models trained with different time window sizes.

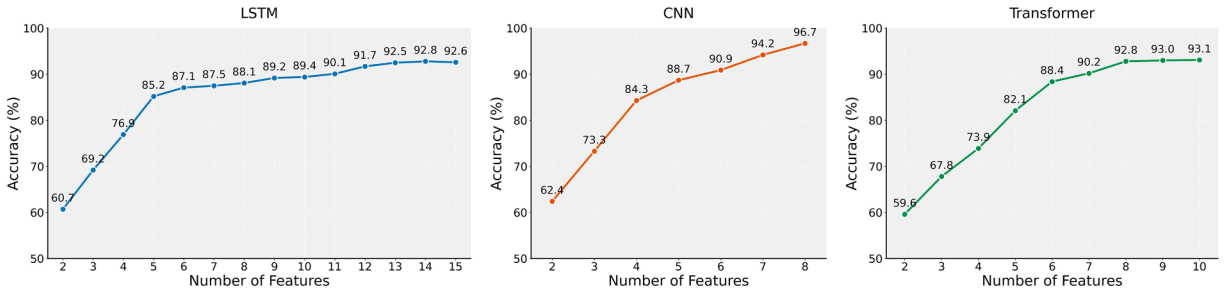

Fig. 13. Model accuracy variation with different numbers of features.

are removed, the model accuracy does not degrade significantly, but further removal of features leads to a notable performance drop. Moreover, after removing the right lane existence (RLaneExist) feature, the model accuracy slightly increases, suggesting that this feature contributes little to the decision process and may even be redundant.

For the CNN model, the original input includes 448 features, among which some are essential for convolutional representation. Due to the large number of features, this experiment mainly focuses on the 8 core vehicle features as discussed in Section 4.7. The results show that CNN exhibits greater sensitivity to these features; removing any features results in a clear drop in model accuracy. The Transformer model demonstrates a similar pattern in the ablation study, where the removal of features significantly affects model performance. These findings suggest that, compared to LSTM, both CNN and Transformer models rely more heavily on core vehicle features and their performance is more sensitive to feature completeness.

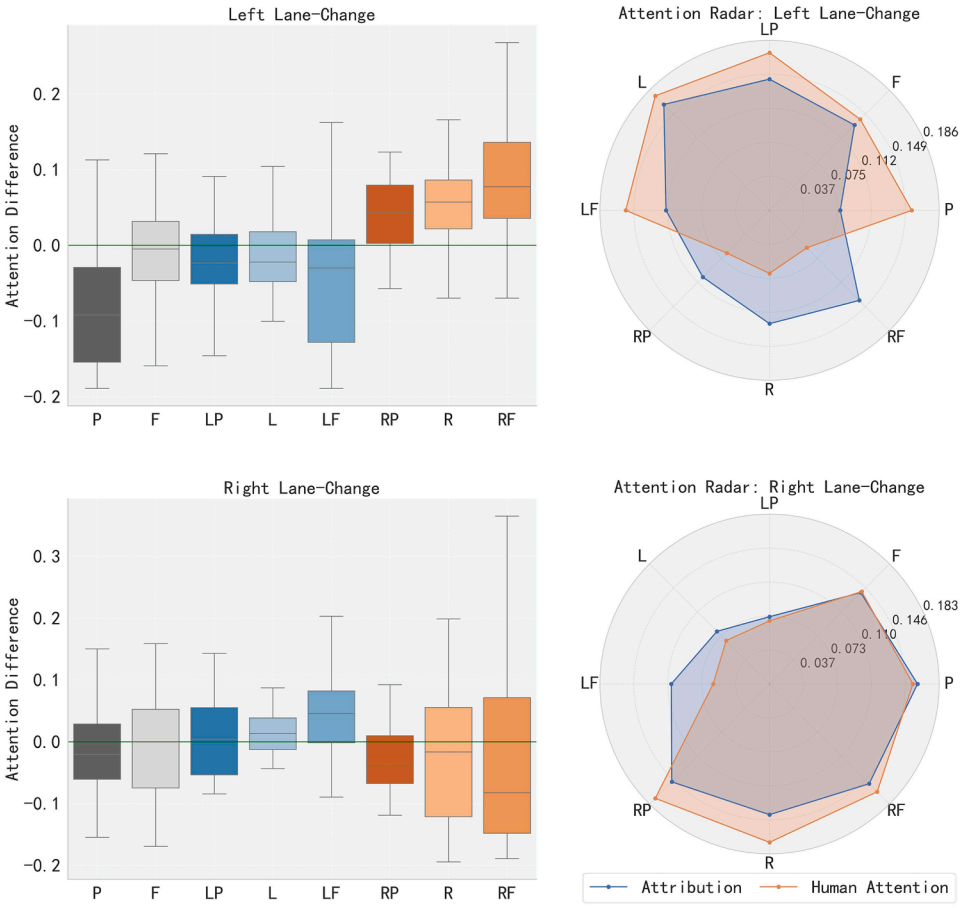

Fig. 14. Comparison of alignment between the decision-making of the LSTM model and human attentional focus.

Table 4
Accuracy gains after applying data balancing technique.

	LSTM	CNN	Transformer
Accurate gain	+0.51	+0.32	+0.46

We further cross-analyze the LSTM model's feature attribution results with the feature selection. The analysis reveals that the attribution score of the RLaneExist feature consistently remains low as discussed in Section 4.7, which aligns with its limited contribution observed in the ablation study. After removing this feature, the accuracy of the LSTM model even improves slightly.

4.10. Impact of data balancing

In this section, we analyze the issue of data imbalance by designing and systematically evaluating a data balancing scheme, rather than simply following existing settings. Specifically, we apply the synthetic minority oversampling technique (SMOTE) as referred to (Ali et al., 2022; Chawla et al., 2002), a neighbor-based synthetic oversampling method, to the lane-change class data in order to optimize the data distribution and retrain the CNN, LSTM, and Transformer models. To quantify the performance improvement brought by data balancing, we calculate the accuracy gain in percentage points.

Table 4 show that, compared to the baseline without data balancing, all models achieve performance improvements, with the LSTM model demonstrating the greatest accuracy gain. This indicates that data balancing techniques are highly practical for improving the accuracy of lane-change models.

On the optimized versions of all three models, we further conduct attribution experiments and evaluate the attribution results using four quantitative attribution metrics. The results shown in Table 5 indicate that our attribution method yields similar results on both the original and balanced models. This suggests that the proposed attribution method exhibits strong robustness and accuracy across different models and data distributions.

Table 5
Attribution metric evaluation for models before and after data balancing training.

	Sen-n ↑	AIC ↓	SIC ↓	MRF ↓
Original LSTM	0.765	0.172	0.186	0.082
Data balanced LSTM	0.757	0.171	0.185	0.081
Original CNN	0.706	0.242	0.251	0.128
Data balanced CNN	0.702	0.244	0.255	0.121
Original Transformer	0.756	0.195	0.203	0.089
Data balanced Transformer	0.751	0.199	0.203	0.091

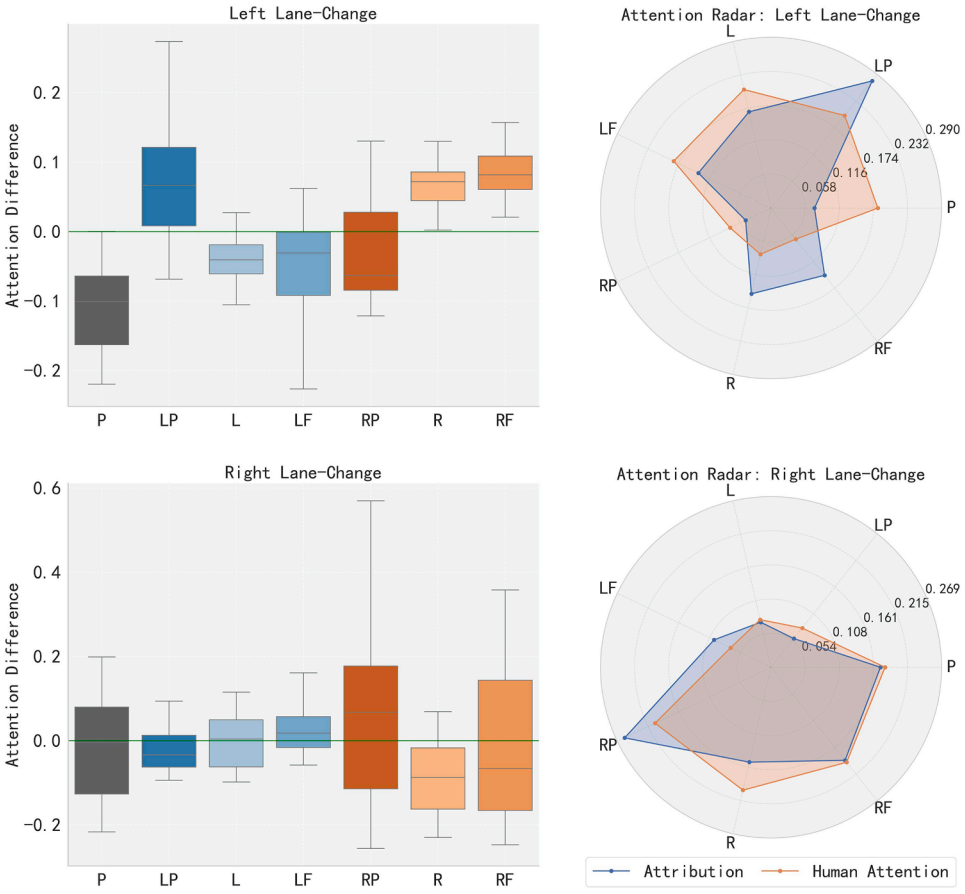

Fig. 15. Comparison of alignment between the decision-making of the CNN model and human attentional focus.

4.11. Human-model decision alignment analysis

This section details a user study designed to assess the degree of alignment between the decision-making processes of deep lane-change models and human intuition. The previous experiments reveal some similarities between the information considered by deep lane-change models and human drivers during decision-making. To further validate this finding, we conduct this experiment.

We randomly selected 50 left lane-change scenarios and 50 right lane-change scenarios from each of the highD and NGSIM datasets. Feature attribution calculations were performed on these scenarios to identify the key features influencing the model's decisions. For the LSTM model, we primarily evaluated the time-to-collision (TTC) with eight surrounding vehicles to analyze the influence of vehicles at different locations. For the CNN and Transformer models, we summed the attribution matrix for each surrounding vehicle to represent its contribution.

To compare model attributions with human judgment, we recruited 67 participants from three universities and two surrounding local communities, ensuring a diverse sample in terms of age (from 19 to 52 years) and driving experience. Notably, participants varied widely in their driving experience, ranging from novices with only a few months behind the wheel to highly experienced drivers with over 20 years of driving. This diversity allowed us to capture a broad spectrum of perspectives and behaviors in the evaluation process. After carefully observing each scenario, participants were asked to select the three most important vehicles they

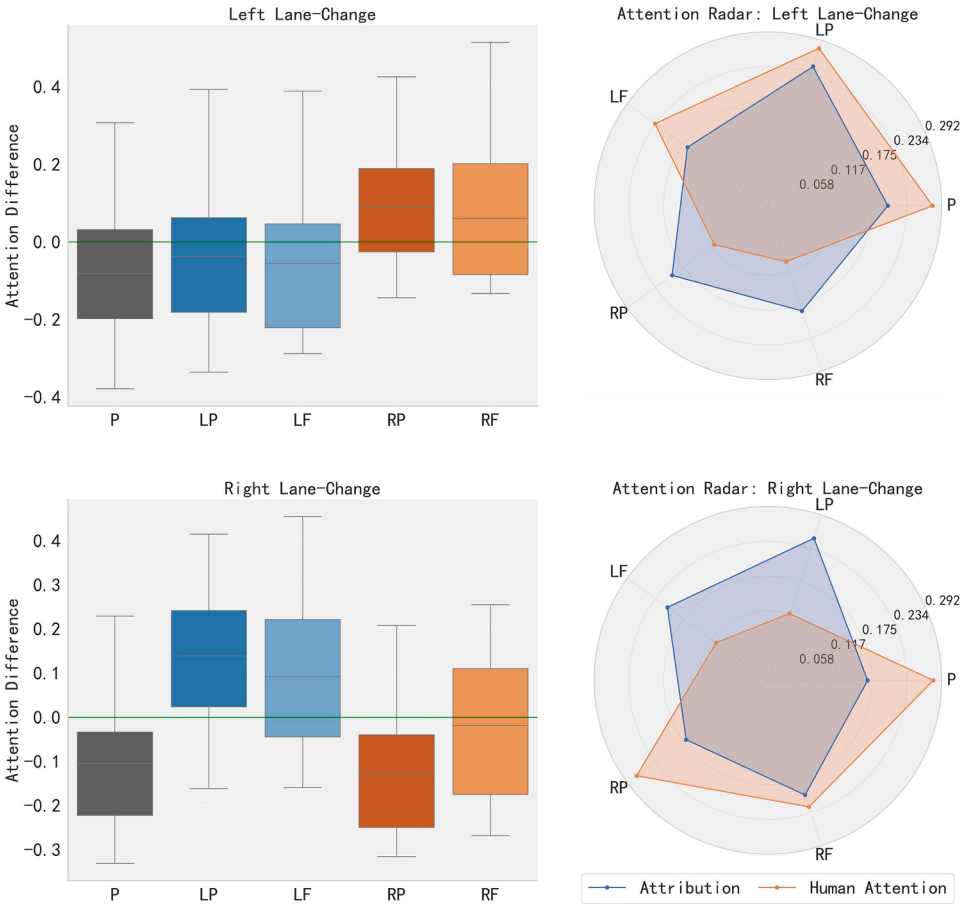

Fig. 16. Comparison of alignment between the decision-making of the Transformer model and human attentional focus.

would consider when making a lane-change decision. Their selections were converted into a binary vector, where 1 stands for a selected vehicle and 0 for an unselected vehicle. The results from all participants were then averaged, normalized, and compared with the normalized attribution results from the deep learning models.

The left side of Fig. 14 presents the results of subtracting human attention from the LSTM model's attribution scores. Values above 0 indicate greater model emphasis on a feature, while values below 0 indicate greater human emphasis. The results reveal some differences in attentional focus, with humans appearing to focus more on information from the target lane of the lane change. For instance, in left lane changes, humans pay more attention to vehicles on the left. The radar charts on the right side of Fig. 14 illustrate the attention distribution of both humans and models, where greater overlap indicates higher alignment. These charts demonstrate a moderate degree of consistency between human and model attention patterns for both left and right lane changes.

Fig. 15 presents the comparison results for the CNN model. For left lane changes, the CNN model tends to focus more heavily on the left preceding vehicle, whereas human drivers distribute their attention more evenly among all vehicles on the left. This indicates that the CNN model may rely on a narrower set of features in certain scenarios. For right lane changes, however, the radar chart shows a closer alignment between the model and human attention patterns, with both giving more balanced consideration to relevant vehicles. Overall, although the CNN model captures some aspects of human intuition, it can still exhibit a more selective focus than human drivers, especially when assessing vehicles in the target lane.

Fig. 16 presents the comparison results for the Transformer model. For left lane changes, the box plots indicate that the Transformer model distributes its attention more evenly across all surrounding vehicles, in contrast to the CNN model and human participants. Humans tend to focus more on vehicles in the target lane, particularly the left preceding vehicle, as reflected in the radar chart. The relatively small variance among the Transformer's attributions suggests that it treats information from all lanes as important when making lane-change decisions. A similar pattern is observed for right lane changes, *i.e.*, the Transformer model continues to allocate attention in a balanced manner among all relevant vehicles, whereas human participants again show a clear preference for vehicles in the target lane. Overall, compared to human intuition, which typically emphasizes the conditions in the target lane, the Transformer model relies on a more holistic consideration of all surrounding vehicles when making lane-change decisions.

These results, derived from different model architectures, demonstrate that deep models have developed decision-making capabilities that align with human intuition to some degree. Both the CNN and LSTM models show attention patterns that are more consistent

with human reasoning. In contrast, the Transformer model exhibits a more balanced attention across multiple surrounding vehicles, reflecting a comprehensive assessment strategy that differs from human behavior.

5. Conclusion

In this work, we propose an attribution method capable of accurately generating decision explanations for various types of lane-change models. The primary challenge in performing attribution calculations for lane-change models lies in effectively representing the transition from the absence to the presence of multi-dimensional features while traversing the complex input feature distribution. Based on this observation, we design a path aggregation attribution method. Rather than searching for a single optimal path, we leverage the exponential family of distributions to generate probabilistic paths. This strategy statistically increases the number of path segments that effectively traverse the input feature distribution. Furthermore, we tailor our approach to generate attribution baselines, *i.e.*, the starting points of attribution paths, based on the input feature types of the lane-change models. These enhancements improve the accuracy of our attribution calculations, achieving state-of-the-art results across various lane-change models and multiple attribution metrics. Moreover, we design two model analysis experiments utilizing the attribution results, offering a practical perspective for understanding deep lane-change models.

Despite the advantages of our method, we identify certain limitations and future directions during its development and experimentation. First, the introduction of probabilistic paths significantly increases computational costs, requiring over 30 times more computation than the original Aumann-Shapley calculation (*e.g.*, $P = 30$). Although GPU parallel programming can mitigate time consumption, the resource demands remain substantial, making our method more suitable for post-hoc validation rather than real-time deployment. This limitation is particularly relevant for commercial autonomous vehicles, where onboard chips have restricted computing power and real-time performance is essential. Therefore, before practical deployment, it is necessary to further investigate whether current or future in-vehicle hardware can support such high computational demands. Developing more efficient algorithms or approximation techniques for generating probability paths remains an open and important direction for enabling the practical application of attribution methods. Second, many lane-change models incorporate controllers, such as MPC following the LSTM model (Wang et al., 2022), but our current experiments simplify the controller component. Extending the attribution method to account for controller operations represents a useful direction for future research. Generating end-to-end explanations that integrate both model predictions and controller behavior would provide a more comprehensive understanding of complex decision-making in driving scenarios. Third, exploring additional state-of-the-art lane-change models and evaluating the effectiveness of attribution methods as well as the reliability of decision-making processes would also constitute a valuable direction for future research. In particular, unifying multiple lane-change datasets and training models with different architectures for subsequent attribution analysis could help us gain a clearer understanding of the implicit relationships between data characteristics and lane-change models.

CRediT authorship contribution statement

Rui Shi: Writing – original draft, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization; **Tianxing Li:** Writing – review & editing, Visualization, Software, Investigation, Funding acquisition; **Yasushi Yamaguchi:** Writing – review & editing, Validation, Resources, Investigation, Funding acquisition; **Liguo Zhang:** Writing – review & editing, Validation, Investigation, Funding acquisition, Formal analysis.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62403017, 62402021, U2233211), Beijing Natural Science Foundation (Grant Nos. 4244088, L243026), and Japan Society for the Promotion of Science (JSPS KAKENHI Grant Nos. 20H04203, 25K03125).

References

Afederal Highway Administration, 2021. Ngsim-next generation simulation https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm.

Ali, Y., Hussain, F., Bliemer, M.C.J., Zheng, Z., Haque, M.M., 2022. Predicting and explaining lane-changing behaviour using machine learning: a comparative study. Transp. Res. Part C Emerg. Technol. 145, 103931. https://doi.org/10.1016/j.trc.2022.103931

Ali, Y., Sharma, A., Zheng, Z., 2025. Empirical research on car-following and lane-changing: Recent developments, emerging vehicle technologies' impact, and future research needs. Transp. Res. Interdiscip. Perspect. 31, 101368. https://doi.org/10.1016/j.trip.2025.101368

Ali, Y., Zheng, Z., Bliemer, M.C.J., 2023. Calibrating lane-changing models: two data-related issues and a general method to extract appropriate data. Transp. Res. Part C Emerg. Technol. 152, 104182. https://doi.org/10.1016/j.trc.2023.104182

Anik, B.M.T.H., Islam, Z., Abdel-Aty, M., 2024. A time-embedded attention-based transformer for crash likelihood prediction at intersections using connected vehicle data. Transp. Res. Part C Emerg. Technol. 169, 104831. https://doi.org/10.1016/j.trc.2024.104831
Aumann, R.J., Shapley, L.S., 1974. Values of Non-Atomic Games. Princeton University Press.
Bassi, P.R., Dertkigil, S.S.J., Cavalli, A., 2024. Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization. Nat. Commun. 15 (1), 291.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/JAIR.953
Chen, H., Covert, I.C., Lundberg, S.M., Lee, S., 2023. Algorithms to estimate Shapley value feature attributions. Nat. Mac. Intell. 5 (6), 590–601. https://doi.org/10.1038/S42256-023-00657-X
Chen, X., Qin, G., Seo, T., Yin, J., Tian, Y., Sun, J., 2024. A macro-micro approach to reconstructing vehicle trajectories on multi-lane freeways with lane changing. Transp. Res. Part C Emerg. Technol. 160, 104534. https://www.sciencedirect.com/science/article/pii/S0968090X2400055X. https://doi.org/10.1016/j.trc.2024.104534
Chen, X.w., Jeong, J.C., 2007. Enhanced recursive feature elimination. In: Int. Conf. Mach. Learn. Appl., pp. 429–435. https://doi.org/10.1109/ICMLA.2007.35
Cheng, S., Wang, Z., Yang, B., Nakano, K., 2024. Convolutional neural network-based lane-change strategy via motion image representation for automated and connected vehicles. IEEE Trans. Neural Netw. Learn. Syst. 35 (9), 12953–12964. https://doi.org/10.1109/TNNLS.2023.3265662
Chormai, P., Herrmann, J., Müller, K.R., Montavon, G., 2024. Disentangled explanations of neural network predictions by finding relevant subspaces. IEEE Trans. Pattern Anal. Mach. Intell. 46 (11), 7283–7299. https://doi.org/10.1109/TPAMI.2024.3388275
Cortes, C., Mohri, M., Rostamizadeh, A., 2012. Algorithms for learning kernels based on centered alignment. J. Mach. Learn. Res. 13, 795–828. https://doi.org/10.5555/2503308.2188413
Deng, H., Zou, N., Du, M., Chen, W., Feng, G., Yang, Z., Li, Z., Zhang, Q., 2024. Unifying fourteen post-hoc attribution methods with taylor interactions. IEEE Trans. Pattern Anal. Mach. Intell. 46 (7), 4625–4640. https://doi.org/10.1109/TPAMI.2024.3358410
Dong, J., Chen, S., Miralinaghi, M., Chen, T., Li, P., Labi, S., 2023. Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems. Transp. Res. Part C Emerg. Technol. 156, 104358. https://doi.org/10.1016/j.trc.2023.104358
Du, G., Zou, Y., Zhang, X., Li, Z., Liu, Q., 2023. Hierarchical motion planning and tracking for autonomous vehicles using global heuristic based potential field and reinforcement learning based predictive control. IEEE Trans. Intell. Transp. Syst. 24 (8), 8304–8323. https://doi.org/10.1109/TITS.2023.3266195
Feng, Y., Hua, W., Sun, Y., 2023. NLE-DM: natural-language explanations for decision making of autonomous driving based on semantic scene understanding. IEEE Trans. Intell. Transp. Syst. 24 (9), 9780–9791. https://doi.org/10.1109/TITS.2023.3273547
Gao, K., Li, X., Chen, B., Hu, L., Liu, J., Du, R., Li, Y., 2023. Dual transformer based prediction for lane change intentions and trajectories in mixed traffic environment. IEEE Trans. Intell. Transp. Syst. 24 (6), 6203–6216. https://doi.org/10.1109/TITS.2023.3248842
Guo, H., Keyvan-Ekbatani, M., Xie, K., 2024. Modeling coupled driving behavior during lane change: A multi-agent transformer reinforcement learning approach. Transp. Res. Part C Emerg. Technol. 165, 104703. https://doi.org/10.1016/j.trc.2024.104703
Gupta, A., Choudhary, P., Parida, M., 2024. Analyzing lane change execution behavior on expressway using an instrumented vehicle: a random effect accelerated failure time approach. IEEE Trans. Intell. Transp. Syst. 25 (11), 18038–18048. https://doi.org/10.1109/TITS.2024.3436569
Janizek, J.D., Dincer, A.B., Celik, S., Chen, H., Chen, W., Naxerova, K., Lee, S.I., 2023. Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models. Nat. Biomed. Eng. 7 (6), 811–829.
Kapishnikov, A., Bolukbasi, T., Viégas, F.B., Terry, M., 2019. XRAI: better attributions through regions. In: Proc. IEEE Int. Conf. Comput. Vis., pp. 4947–4956. https://doi.org/10.1109/ICCV.2019.00505
Kim, C., Gadgil, S.U., DeGrave, A.J., Omiye, J.A., Cai, Z.R., Daneshjou, R., Lee, S.I., 2024. Transparent medical image AI via an image-text foundation model grounded in medical literature. Nat. Med. , 1–12.
Kornblith, S., Norouzi, M., Lee, H., Hinton, G.E., 2019. Similarity of neural network representations revisited. In: Proc. Int. Conf. Mach. Learn.. Vol. 97, pp. 3519–3529.
Krajewski, R., Bock, J., Kloeker, L., Eckstein, L., 2018. The highD dataset: a drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In: Int. Conf. Intell. Transp. Syst., pp. 2118–2125. https://doi.org/10.1109/ITSC.2018.8569552
Li, D., Deng, H., Yu, T., Zhang, L., 2025a. Multi-vehicle cooperative localization using a TOA-based simulated annealing extended kalman filter in urban canyons. IEEE Internet Things J. 12 (13), 22832–22846. https://doi.org/10.1109/JIOT.2025.3550553
Li, J., Zhang, H., Liang, S., Dai, P., Cao, X., 2023a. Privacy-enhancing face obfuscation guided by semantic-aware attribution maps. IEEE Trans. Inf. Forensics Security 18, 3632–3646. https://doi.org/10.1109/TIFS.2023.3282384
Li, M., Sun, H., Cui, Z., Huang, Y., Chen, H., 2024a. Expected integral discrete gradient: Diagnosing autonomous driving model. IEEE Trans. Veh. Technol. 73 (12), 18198–18207. https://doi.org/10.1109/TVT.2024.3436614
Li, M., Sun, H., Huang, Y., Chen, H., 2024b. SVCE: shapley value guided counterfactual explanation for machine learning-based autonomous driving. IEEE Trans. Intell. Transp. Syst. 25 (10), 14905–14916. https://doi.org/10.1109/TITS.2024.3393634
Li, M., Wang, Y., Sun, H., Cui, Z., Huang, Y., Chen, H., 2023b. Explaining a machine-learning lane change model with maximum entropy Shapley values. IEEE Trans. Intell. Veh. 8 (6), 3620–3628. https://doi.org/10.1109/TIV.2023.3266196
Li, T., Shi, R., Li, Z., Kanai, T., Zhu, Q., 2024c. Efficient deformation learning of varied garments with a structure-preserving multilevel framework. Proc. ACM Comput. Graph. Interact. Tech. 7 (1), 1–19. https://doi.org/10.1145/3651286
Li, T., Shi, R., Zhu, Q., Zhang, L., Kanai, T., 2025b. Spectrum-enhanced graph attention network for garment mesh deformation. IEEE Trans. Pattern Anal. Mach. Intell. , 1–18. https://doi.org/10.1109/TPAMI.2025.3570523
Li, Y., Jiang, Y., Wu, X., 2025c. TrajPT: a trajectory data-based pre-trained transformer model for learning multi-vehicle interactions. Transp. Res. Part C Emerg. Technol. 171, 105013. https://doi.org/10.1016/j.trc.2025.105013
Li, Y., Liang, H., Zheng, L., 2025d. WB-LRP: layer-wise relevance propagation with weight-dependent baseline. Pattern Recognit. 158, 110956. https://doi.org/10.1016/J.PATCOG.2024.110956
Li, Y., Shi, T., Chen, Z., Zhang, L., Xie, W., 2024d. GT-CAM: game theory based class activation map for GCN. IEEE Trans. Pattern Anal. Mach. Intell. 46 (12), 8806–8819. https://doi.org/10.1109/TPAMI.2024.3413026
Liang, Y., Ding, F., Huang, G., Zhao, Z., 2023. Deep trip generation with graph neural networks for bike sharing system expansion. Transp. Res. Part C Emerg. Technol. 154, 104241. https://doi.org/10.1016/j.trc.2023.104241
Lundberg, S.M., Lee, S., 2017. A unified approach to interpreting model predictions. In: Proc. Adv. Neural Inf. Process. Syst., pp. 4765–4774.
Ma, H., Qian, C., Li, L., manda, H., Qu, X., Ran, B., 2025. A novel 2d motion planning method for vehicles considering the impact of lane configurations. Transp. Res. Part C Emerg. Technol. 177, 105186. https://www.sciencedirect.com/science/article/pii/S0968090X25001901. https://doi.org/10.1016/j.trc.2025.105186
Mamalakis, A., Barnes, E.A., Ebert-Uphoff, I., 2023. Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience. Artif. Intell. Earth Syst. 2 (1), e220058. https://doi.org/10.1175/AIES-D-22-0058.1
Oseni, A., Moustafa, N., Creech, G., Sohrabi, N., Strelzoff, A., Tari, Z., Linkov, I., 2023. An explainable deep learning framework for resilient intrusion detection in IoT-enabled transportation networks. IEEE Trans. Intell. Veh. 24 (1), 1000–1014. https://doi.org/10.1109/TITS.2022.3188671
Schulz, K., Sixt, L., Tombari, F., Landgraf, T., 2020. Restricting the flow: information bottlenecks for attribution. In: Proc. Int. Conf. Learn. Representations. https://openreview.net/forum?id=S1xWh1rYwB.
Shi, R., Li, T., Yamaguchi, Y., Zhang, L., 2025a. Exploring decision shifts in autonomous driving with attribution-guided visualization. IEEE Trans. Intell. Transp. Syst. 26 (3), 4165–4177. https://doi.org/10.1109/TITS.2024.3513400
Shi, R., Li, T., Yamaguchi, Y., Zhang, L., 2025b. Traffic scene-informed attribution of autonomous driving decisions. IEEE Trans. Intell. Transp. Syst. ,26 (7), 9175–9186. https://doi.org/10.1109/TITS.2025.3547879
Shi, R., Li, T., Zhang, L., Yamaguchi, Y., 2024. Visualization comparison of vision transformers and convolutional neural networks. IEEE Trans. Multim. 26, 2327–2339. https://doi.org/10.1109/TMM.2023.3294805

Singla, V., Sandoval-Segura, P., Goldblum, M., Geiping, J., Goldstein, T., 2023. A simple and efficient baseline for data attribution on images. In: Proc. Adv. Neural Inf. Process. Syst. Work. https://openreview.net/forum?id=WGjQt3aDn7.
Su, S., Ju, X., Xu, C., Dai, Y., 2024. Collaborative motion planning based on the improved ant colony algorithm for multiple autonomous vehicles. IEEE Trans. Intell. Transp. Syst. 25 (3), 2792–2802. https://doi.org/10.1109/TITS.2023.3250756
Sun, J., Yang, H., 2024. Learning two-dimensional merging behaviour from vehicle trajectories with imitation learning. Transp. Res. Part C Emerg. Technol. 160, 104530. https://doi.org/10.1016/j.trc.2024.104530
Sun, K., Gong, S., Zhou, Y., Chen, Z., Zhao, X., Wu, X., 2024. A multi-vehicle cooperative control scheme in mitigating traffic oscillation with smooth tracking-objective switching for a single-vehicle lane change scenario. Transp. Res. Part C Emerg. Technol. 159, 104487. https://www.sciencedirect.com/science/article/pii/S0968090X24000081. https://doi.org/10.1016/j.trc.2024.104487
Wang, H., Lu, B., Li, J., Liu, T., Xing, Y., Lv, C., Cao, D., Li, J., Zhang, J., Hashemi, E., 2022. Risk assessment and mitigation in local path planning for autonomous vehicles with LSTM based predictive model. IEEE Trans. Autom. Sci. Eng. 19 (4), 2738–2749. https://doi.org/10.1109/TASE.2021.3075773
Wang, Y., Gloudemans, D., Ji, J., Teoh, Z.N., Liu, L., Zachár, G., Barbour, W., Work, D., 2024. Automatic vehicle trajectory data reconstruction at scale. Transp. Res. Part C Emerg. Technol. 160, 104520. https://doi.org/10.1016/j.trc.2024.104520
Xing, Y., Wu, Y., Wang, H., Wang, L., Li, L., Peng, Y., 2025. Failed lane-changing detection and prediction using naturalistic vehicle trajectories. Transp. Res. Part C Emerg. Technol. 170, 104939. https://www.sciencedirect.com/science/article/pii/S0968090X24004601. https://doi.org/10.1016/j.trc.2024.104939
Xue, Y., Wang, C., Ding, C., Yu, B., Cui, S., 2024. Observer-based event-triggered adaptive platooning control for autonomous vehicles with motion uncertainties. Transp. Res. Part C Emerg. Technol. 159, 104462. https://doi.org/10.1016/j.trc.2023.104462
Yang, R., Wang, B., Bilgic, M., 2023. IDGI: a framework to eliminate explanation noise from integrated gradients. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 23725–23734. https://doi.org/10.1109/CVPR52729.2023.02272
Yao, R., Sun, X., 2025. Hierarchical prediction uncertainty-aware motion planning for autonomous driving in lane-changing scenarios. Transp. Res. Part C Emerg. Technol. 171, 104962. https://www.sciencedirect.com/science/article/pii/S0968090X24004832. https://doi.org/10.1016/j.trc.2024.104962
Zhan, J., Zhang, L., Qiao, J., 2025. Boundary consensus of networked hyperbolic systems of conservation laws. IEEE Trans. Autom. Control , 1–16. https://doi.org/10.1109/TAC.2025.3529281
Zhang, B., Zheng, W., Zhou, J., Lu, J., 2024. Path choice matters for clear attributions in path methods. In: Proc. Int. Conf. Learn. Representations. https://openreview.net/forum?id=gzYgsZgwXa.
Zhang, K., Li, L., 2022. Explainable multimodal trajectory prediction using attention models. Transp. Res. Part C Emerg. Technol. 143, 103829. https://www.sciencedirect.com/science/article/pii/S0968090X22002509. https://doi.org/10.1016/j.trc.2022.103829
Zhang, Q., Cheng, X., Chen, Y., Rao, Z., 2023a. Quantifying the knowledge in a DNN to explain knowledge distillation for classification. IEEE Trans. Pattern Anal. Mach. Intell. 45 (4), 5099–5113. https://doi.org/10.1109/TPAMI.2022.3200344
Zhang, Y., Xu, Q., Wang, J., Wu, K., Zheng, Z., Lu, K., 2023b. A learning-based discretionary lane-change decision-making model with driving style awareness. IEEE Trans. Intell. Transp. Syst. 24 (1), 68–78. https://doi.org/10.1109/TITS.2022.3217673