# Understanding Decision-Making of Autonomous Driving via Semantic Attribution

Rui Shi, Tianxing Li, Yasushi Yamaguchi, *Member, IEEE*, and Liguo Zhang, *Member, IEEE*

***Abstract*— Understanding decision-making in autonomous driving models is essential for real-world applications. Attribution explanation is a primary research direction for interpreting neural network decisions. However, in the context of autonomous driving, numerical attributions fail to interpret the complex semantic information and often result in explanations that are difficult to understand. This paper introduces a novel semantic attribution approach that both identifies where important features appear and provides intuitive information about what they represent. To establish the semantic correspondences for attributions, we propose an interpreting framework that integrates unsupervised differentiable semantic representations with the attribution computational model. To further enhance the accuracy of the attribution computation while ensuring strong semantic correspondence, we design a Semantic-Informed Aumann-Shapley (SIAS) method, which defines a novel integration path solution using constraints from semantic scores and discrete gradients. Extensive experiments confirm that our method outperforms state-of-the-art explanation techniques both qualitatively and quantitatively in autonomous driving scenarios.**

***Index Terms*— Autonomous driving, attribution explanation, semantic representation, Shapley value.**

## I. Introduction

AUTONOMOUS driving has the potential to significantly reduce traffic accidents and enhance road safety, capturing the interest of the transportation, robotics, and artificial intelligence research communities [1]. Recent advancements in artificial intelligence have driven remarkable progress across a wide array of tasks related to autonomous driving. There are various types of deep neural networks (DNNs) have been developed, each tailored for specific applications *e.g.*, traffic object detection [2], [3], scene understanding [4], [5], vehicle localization [6], [7], motion planning [8], [9], trajectory tracking [10], [11], and end-to-end decision-making [12], [13].

Received 18 May 2024; revised 20 July 2024 and 11 September 2024; accepted 15 October 2024. Date of publication 29 October 2024; date of current version 9 January 2025. This work was supported in part by Beijing Natural Science Foundation under Grant 4244088 and Grant L243026; in part by the National Natural Science Foundation of China under Grant 62403017, Grant 62402021, and Grant U2233211; and in part by Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 20H04203. The Associate Editor for this article was S. Santini. *(Corresponding author: Tianxing Li.)*

Rui Shi and Liguo Zhang are with the School of Information Science and Technology, Beijing University of Technology, Beijing 100124, China (e-mail: ruishi@bjut.edu.cn; zhangliguo@bjut.edu.cn).

Tianxing Li is with the College of Computer Science, Beijing University of Technology, Beijing 100124, China (e-mail: litianxing@bjut.edu.cn).

Yasushi Yamaguchi is with the Department of General Systems Studies, The University of Tokyo, Tokyo 153-8902, Japan (e-mail: yama@graco.c.u-tokyo.ac.jp).

Digital Object Identifier 10.1109/TITS.2024.3483810

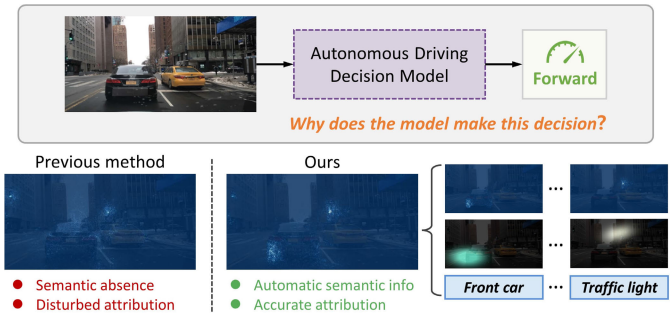



Fig. 1. Attribution results with and without semantic information. The explanation result of [14] is shown on the left. The results generated by our attribution method are shown on the right. Attribution results that incorporate semantic information are more comprehensible and help avoid potential misunderstandings due to noise disturbance.

Although DNNs are essential for advancing autonomous driving technologies, they present a significant challenge: decision-making processes lack interpretability. This black-box nature stems from the fact that DNN operations involve nested sequences of complex nonlinear functions. The difficulty in understanding how inputs affect driving decisions presents a substantial barrier to DNN deployment in real-world scenarios, particularly in dynamic and unpredictable driving environments where trust and reliability are paramount. To address this challenge, researchers have developed various attribution methods [15], [16]. These methods aim to clarify decision-making by quantifying the influence of specific input features on the decisions, thus enhancing the transparency and trustworthiness of autonomous driving models.

Attribution methods are versatile and can be applied to diverse application scenarios; however, decision-making of autonomous driving models presents a unique challenge due to the random distribution of numerous object semantics within the environment. Previous attribution methods primarily calculate feature contributions but fall short in establishing causal relationships between object semantics and decisions. This means that explanations can only provide limited insights, indicating only "where" important features are located, without clarifying "what" features represent, as depicted in the bottom-left corner of Fig. 1. Furthermore, the complexity of traffic scenarios often leads to attribution results being disturbed by noise, making it more challenging to accurately and directly understand the semantic meanings of features. Consequently, these partial insights place the burden of interpreting the model reasoning squarely on users.

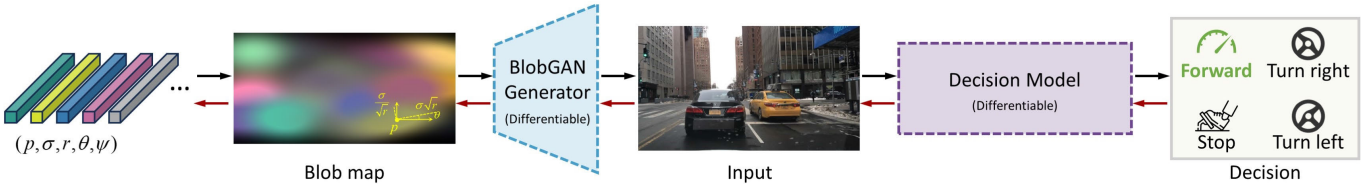



Fig. 2. Fully differentiable autonomous driving decision-making process. Starting from blob information, an input image is generated, which then passes through the autonomous driving model to obtain the decision. Additionally, the entire process supports backpropagation. Black arrows indicate forward propagation and red arrows indicate backpropagation.

In this context, our goal is to answer both the "where" and "what" questions regarding the model decision during driving. This involves not only providing precise attribution but also automatically highlighting the distinct semantics, as demonstrated in the bottom-right corner of Fig. 1. The intrinsic value of incorporating semantics into attribution explanations lies in transcending mere heatmaps to create semantic regions that are readily identifiable by humans. This enhancement improves the utility of interpretation tools, enabling researchers and engineers to comprehend the explanation results on a more tangible and recognizable level.

Given the semantic complexity of driving scenarios, labeling semantics manually is impractical. This challenge inspires us to employ unsupervised learning techniques, specifically utilizing a spatial-disentangled semantic representation approach BlobGAN [17], to train a generator that represents semantics as blobs from extensive unlabeled driving data. Subsequently, we introduce the Aumann-Shapley (AS) method [18], adapted for infinite games, as our attribution computational model. The introduction of AS method is facilitated by the differentiability of both the blob generator and the autonomous driving model, enabling the discrete-gradient-based AS method to trace decisions from output back through image space to blob information for attribution calculation (the backpropagation indicated by the red arrows shown in Fig. 2). Our interpreting framework, which integrates unsupervised differentiable semantic representations with the attribution computational model, not only quantifies the influences of input features but also establishes semantic-level correspondences, offering readily comprehensible explanations for driving decisions.

Despite establishing semantic correspondences for attributions, the random distribution of semantics in driving scenarios can still compromise the accuracy of attribution computation. To solve this problem, we design a semantic-informed method to enhance the AS computational model. Theoretically, the AS computation is a path integral where the choice of the integration path should ideally not impact the results. However, our empirical investigations using the BDD100k [19] and BDD-OIA [20] datasets have shown that in practical applications of autonomous driving models, the choice of integration path can significantly impact attribution accuracy. This effect arises from the discretization of gradients and the use of integral approximation methods, and is exacerbated when the integration paths ignore semantic information. With this observation, our intuition is that in a majority of perceptually unconstrained traffic scenarios, the semantic information is coherent; therefore, the integration path should be designed to avoid disruption of coherence within individual semantic regions and to exclude anomalous samples. With this idea, we propose the Semantic-Informed Aumann-Shapley (SIAS) attribution method.

Our SIAS attribution method introduces a novel integration path that progresses from the baseline to the input. At each step, the method adaptively selects features based on the constraints that consider both the influence and the semantics of the features. Unlike previous methods that focus solely on pixel attributions, we extend our approach to include blob attributions, which we refer to as feature attributions. As the magnitude of specific features (pixels in image space, blob feature parameters in blob generator) reaches equivalence with those in the input being interpreted, they are no longer candidates for selection. This approach ensures that the attributions maintain strong semantic correspondences and improve in accuracy throughout path integral. Both qualitative and quantitative experiments conducted on large-scale datasets highlight the advantages of SIAS compared to previous attribution explanation methods. Our main contributions are as follows:

- We propose a framework that simultaneously addresses the "where" and "what" questions of model inference by combining unsupervised spatial-disentangled learning with the AS-based attribution method. This integration can autonomously generate semantically equipped attribution explanations, thereby enhancing the comprehensibility of driving decision attributions.
- We design Semantic-Informed Aumann-Shapley (SIAS), an attribution method that selects an integration path by incorporating semantic blobs and differential constraints. We validate our proposal on two datasets using multiple quantitative and qualitative metrics, demonstrating its ability to accurately capture scene semantics and interpret decision-making.

## II. Related Work

### A. Attribution Methods for Autonomous Driving

The importance of explainability in autonomous driving has received increasing attention by researchers [21], [22]. Recent literature categorizes attribution interpreting methods into those that perturb the input [23], [24], rely on backpropagation [14], [16], [25], [26], [27], [28], and hybrid techniques that combine these methods [29], [30]. Our work enhances the Aumann-Shapley values, a key technique within the backpropagation category, contributing to advancements in this field.

*Perturbation-based methods* modify input features and observe output changes. Sacha et al. [23] utilized semantic segmentation to adaptably identify key input features. Shrikumar et al. [24] created DeepLIFT, which measures neuron activation against a reference state to trace connections between neurons. Despite their flexibility, they often suffer from inefficiency due to the necessity of multiple iterations or operations for each input.

*Backpropagation-based methods* that use the accessible gradients with respect to input features to generate attributions are the most commonly employed technique for autonomous driving models. GradShap [31] is applied to scenarios such as traffic object detection [25] and lane change predictions [16], effectively showing the rationale behind specific autonomous driving decisions. Shi et al. [14] explored the impact of the baseline on Aumann-Shapley attributions and introduced an optimization-based method for generating baselines. Chen et al. [26] proposed PropShapley that constructs a Shapley-value propagation model to facilitate attribution computation in DNNs across different modalities. Bojarski et al. [27] defined a propagation rule (VisualBackProp) similar to deconvolution to visualize attributions in autonomous driving scenarios. Additionally, several DNN attribution methods originally developed for specific fields like biology, economics, and psychology [28], [32], [33], [34], [35] could be adapted to enhance autonomous driving models with slight modifications.

*Hybrid attribution methods* merge techniques like backpropagation and attribution baselines, exemplified by the layer-wise relevance propagation optimization [29] and the attribution aggregating [30]. Although previous methods successfully establish input-output causality, they do not include semantic information in attribution explanations. This limitation can diminish the clarity and utility of attributions in autonomous driving, especially in complex traffic scenarios where a deep understanding of semantic details is essential.

### B. Explainable Autonomous Driving Methods

Building interpretable models of autonomous driving through *attention mechanisms* [36] have been explored in various driving contexts, including object detection [37], [38], [39], motion forecasting [40], [41], [42], driver attention prediction [43], and recent end-to-end models [44], [45], [46], [47]. Attention models of growing popularity produce heatmap explanations that closely resemble those generated by attribution methods. However, there are two key differences between these methods. Firstly, attribution methods come up with theoretical guarantees, such as the axiom constraints of Aumann-Shapley attributions; these are foundational principles integrated during the development of the attribution computational models. Furthermore, attribution methods are highly versatile and can be applied across a diverse range of network structures. In contrast, attention models are generally restricted to specific network structures, with a primary focus on self-attention architectures.

There are explanation methods that employ generative techniques, such as generative adversarial networks (GANs), to create images that help understand autonomous driving decisions by contrasting generated images with the original input [48]. These methods often utilize GAN models inspired by the StyleGAN architecture [49], [50], with examples including OCTET [51] and SAFE [52], [53]. Although our method similarly employs a generative model, it diverges significantly in its application: we generate attribution explanations that directly relate to the decision-making process, whereas the aforementioned methods primarily produce driving scene images that lack explicit relevance to decision-making.

### C. Datasets for Autonomous Driving

Numerous datasets can be used for autonomous driving research, such as nuScenes [54], HDD [55], CityScapes [56], Apolloscape [57], Oxford RobotCar [58], and BDD100K [19]. BDD100K, one of the largest driving video datasets, includes 100k videos that span a wide array of tasks critical to autonomous driving technologies like object and lane detection. This extensive collection not only supports a broad range of computer vision tasks but also facilitates advanced research and development in automated driving systems. Its subset, BDD-OIA [20], selected for its richness in pedestrian, bicycle, and vehicle presence, provides ground truth for four driving decisions and 21 explanation annotations. The selection of BDD datasets is primarily due to their extensive task labels and annotations conducive to interpretability studies. The value of BDD datasets, particularly with BDD-OIA's detailed annotations, extends typical research applications. They are useful in conducting interpretability experiments, which are crucial for validating the decisions made by autonomous driving models. Additionally, the rich annotations facilitate the design of user studies and the creation of robust validation sets, ensuring that the models developed are not only effective but also transparent and understandable.

## III. Semantic Attribution Integration

In this section, we present our framework that integrates attribution with semantics to gain insight into the decision-making process of autonomous driving models. For attribution calculation, we employ the Aumann-Shapley method [18], which relies on discrete gradients. By integrating this method with differentiable semantic representations, we achieve the generation of semantic attributions, providing deeper insights into model decisions.

For semantic representation, we employ BlobGAN [17] to capture the semantic information of traffic scenes in an unsupervised manner. This GAN model encodes traffic scenes into semantic blobs, where each blob represents a specific scene or object. These semantic blobs can then be used to reconstruct the original scenes. Additionally, since BlobGAN is differentiable, it facilitates easy access to discrete gradients. As shown in Fig. 2, the forward propagation in our decision-making process proceeds from blob parameters to the decision, while backward propagation traces from the decision back to the image space and ultimately to the blob parameters.

Regarding the BlobGAN, each traffic scene image is encoded into multiple semantic blobs. Each blob is represented as an ellipse defined by its center coordinates $p \in [0, 1]^2$, scale $\sigma \in \mathbb{R}$, aspect ratio $r \in \mathbb{R}$, and rotation angle $\theta \in [-\pi, \pi]$. These parameters capture the size and orientation of objects within the scene. Additionally, each blob is characterized by structural and style features $\psi \in \mathbb{R}^{d_{\text{fea}}}$, which determine the finer details of the object. For instance, a blob representing a traffic light can alter its shape from circular to square and its color from green to red by adjusting the structural and style features $\psi$.

The key element of generating the blob map is the blob score $S$ which represents the semantic distribution. We employ the squared Mahalanobis distance as the metric for $S$, aligning with the approach originally used in [17]. The formulation of the blob score is specified as follows:

$$S = \text{sigmoid}(\sigma - (\Delta p)^T (R\Sigma R)^{-1}(\Delta p)), \tag{1a}$$

$$\Delta p = p_{grid} - p, \tag{1b}$$

$$\Sigma = \begin{bmatrix} r & 0 \\ 0 & \frac{1}{r} \end{bmatrix}, \tag{1c}$$

where $\Sigma$ represents an ellipse influenced by the aspect ratio $r$, and may require additional scaling to adjust blob edge in programming implementation. $p_{grid} \in \{(w/W, h/H)\}_{w,h}^{W,H}$ represents the coordinates on a grid normalized by the image dimensions $W$ (width) and $H$ (height). $p$ stands for the center coordinate of a blob ellipse. $\sigma$ denotes the scale of a blob ellipse. $R$ is the 2D rotation matrix corresponding to the rotation angle $\theta$. The function sigmoid is used for smoothing the transition of the blob score.

Up to this point, blobs contain only implicit semantic information, not direct natural language descriptions, *i.e.*, we still cannot obtain language-level semantic results. To enhance semantic interpretation, we further assign labels to each blob by examining how changes in pixels intersect with semantic segmentation labels from the BDD dataset, such as "traffic light," "car," and "building." Additionally, we manually append directional labels like "front," "left," and so forth. While this process still requires manual input, the task of annotating several blobs is relatively minor compared to the extensive effort needed to label a large dataset. The blobs learned through unsupervised learning on the large-scale BDD dataset possess semantic representation capabilities. By encoding images into these blobs, an automatic correspondence between the image regions and blob semantics can be established. Importantly, the transformation process from image to blob is differentiable, allowing for the direct application of attribution techniques. This capability facilitates the establishment of a causal relationship between specific image regions and the corresponding blobs. As a result, it becomes feasible to autonomously assign natural language-level semantic information to the attribution results, enhancing the interpretability and practical utility of the model outputs in real-world applications.

For the attribution computation model, the detailed discussion will be introduced in Sec. IV. For now, it suffices to say that, accurate attributions can be calculated with discrete gradients based on the AS method. The characteristics of our method are visualized in Fig. 3. Initially, the attribution calculations to the image space may seem to lack explicit semantic connections, as shown in the top-middle part of Fig. 3. The attribution calculation typically yields only a heatmap, which places the burden of deciphering specific meanings on the user, which can be challenging to interpret. For instance, the traffic light and the building on the left side of this image overlap, discerning the semantic information solely from the heatmap becomes unfeasible. However, by tracing these calculations back to the blob parameters, we can equip attributions with specific and clear semantics. Moreover, attributions can be further examined statistically by employing a bar chart or other statistical techniques. Semantic attribution enables a deeper and more structured analysis of how different input features contribute to decision-making.

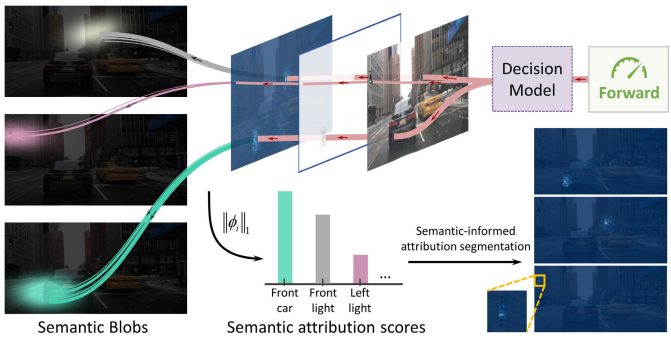



Fig. 3. A brief overview of the methodological contributions of this work. Our attribution method can further backpropagate to the blob parameter after backtracking to the image space, equipping attributions with explicit semantic information.

## IV. Semantic-Informed Aumann-Shapley Attribution Method

In this section, we introduce the adaptation of Aumann-Shapley values to DNNs and our semantic-informed attribution computation method.

### A. Aumann-Shapley Values in DNNs

In game theory, the Aumann-Shapley value is used to quantify the contribution of each participant to a specific outcome when all participants are involved in the game. The fundamental concept can be adapted to DNNs, allowing for the calculation of neuron contributions. Consider an input $x$ and a corresponding baseline $\bar{x}$, where $\bar{x}$ represents the scenario with missing information. We define an integration path $\mu(t)$ between $x$ and $\bar{x}$, parameterized by $t \in [0, 1]$, where $\mu_i(0) = \bar{x}_i$ and $\mu_i(1) = x_i$ represent the values of the $i$-th feature at the beginning and end of the path, respectively. For the original AS computation, the path is defined as a straight-line path given by $\mu(t) = (1-t)\bar{x} + tx$. Note that this path definition does not incorporate any semantic or scenario-related information. In the context of the autonomous driving model, the function $f$ represents the model itself, while $f^d$ denotes the output decision $d$ generated by the model. Given the output $f^d$, the initial idea of the Aumann-Shapley method

involves assessing the marginal contributions of features:

$$\phi_i = \int_{t=0}^{1} \left( f^d\left(\mu(t) + \Delta x_i\right) - f^d\left(\mu(t)\right) \right) dt, \tag{2}$$

where $\Delta x_i = x_i - \bar{x}_i$ denotes the change in feature value along the path. This mechanism enables the measurement of how modifications in feature values influence the output. To further illustrate the sensitivity of the output to changes in input, $f^d\left(\mu(t) + \Delta x_i\right)$ undergoes a Taylor series expansion:

$$f^d\left(\mu(t)+\Delta x_i\right) = f^d\left(\mu(t)\right) + \Delta x_i \frac{\partial f^d\left(\mu(t)\right)}{\partial x_i} + O\left[(\Delta x_i)^2\right], \tag{3}$$

where the remainder term $O\left[(\Delta x_i)^2\right]$ is ignored in practice. Then, Eq. (3) can be incorporated into Eq. (2) to yield a refined expression of the Aumann-Shapley value:

$$\phi_i = \Delta x_i \int_{t=0}^{1} \frac{\partial f^d\left(\mu(t)\right)}{\partial x_i} dt. \tag{4}$$

where the Aumann-Shapley value $\phi_i(\mu)$ is the gradient integral along the path described by $\mu(t)$. However, directly performing integrals in DNNs is impractical due to the discrete nature of data inputs and the discontinuities introduced by activation functions. Moreover, the high dimensionality and complexity of networks make such computations computationally intensive. Therefore, we discretize the continuous integral using Gauss-Legendre quadrature as follows:

$$\phi_i = \Delta x_i \sum_{k=1}^{K} \frac{1}{\left(1-\xi_k^2\right)\left[P_K'(\xi_k)\right]^2} \frac{\partial f^d\left(\mu(\xi_k)\right)}{\partial x_i}, \tag{5}$$

where $K$ is the number of sample points used in the Gauss-Legendre quadrature, each contributing to the discrete approximation of the integral that quantifies the influence of the input feature on the output. The term $\xi_k$ represents the quadrature point of the $k$-th Legendre polynomial. $P_K'$ is the derivative of Legendre polynomials at the sample point. $\mu(\xi_k)$ is the evaluation of the path $\mu$ at these Gauss-Legendre sample points.

### B. SIAS Computational Model

Despite the Aumann-Shapley values offering ideal theoretical properties for measuring feature attributions, its application to DNNs often yields inaccurate attributions in regions independent of decision-making. The choice of integration path $\mu(\cdot)$ significantly influences the attributions results. An ideal path should effectively identify regions that influence decision-making while distinguishing among semantic regions, ensuring that the movement of individual semantic regions along the path is gradual. However, the commonly used straight-line path in original Aumann-Shapley values, which uniformly interpolates input features from a baseline to the input, fails to achieve this. It treats all regions equally, without considering the actual semantic distribution, thus breaking the coherence within individual semantic regions during the path integral computation. Consequently, this approach allows features that do not really contribute to the final output to mistakenly receive non-zero attribution scores.

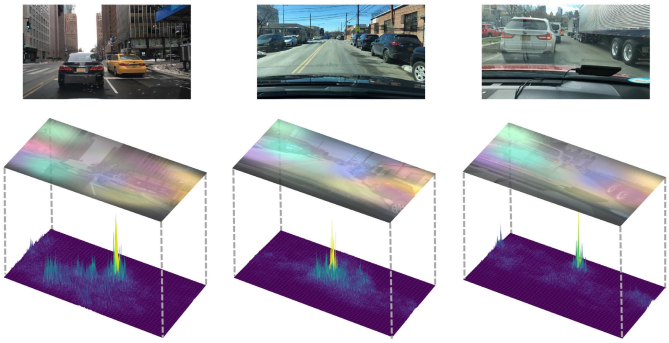


Fig. 4. Semantic-informed integration path selection. The top row shows the original images, and the middle row displays blob maps. The 3D images below are determined by combining semantic blobs with gradient information. During the path integral of the attribution calculation, the regions corresponding to lower values in the 3D image are first modified from the baseline to the input features, followed by modifications to the regions with higher values. This process can be viewed as a gradual transition from the outside (lower values) to the inside (higher values) of individual semantic regions, with the path being sequentially ordered within each semantic region. The peaks on the left and right images are at traffic light regions. These images are best viewed on screen.

To address this issue, we propose the Semantic-Informed Aumann-Shapley (SIAS) attribution computation method. SIAS defines a path selection from the baseline to the input, selectively modifying features that satisfy both semantic region constraints and gradient requirements, as shown in Fig. 4. In this configuration, the pixel values do not change gradually as in the original AS path; instead, they transition directly from the selected baseline values to the input values. By doing this at each step of the modification process, SIAS effectively minimizes the attributions in semantically irrelevant regions, thereby refining the accuracy of the attribution. Specifically, we first define constraints that include semantic correlations:

$$\mathcal{L}_{sc} = \left\| \int_{t=0}^{1} \log(1+S) \left| \frac{\partial f^d\left(\mu(t)\right)}{\partial\left(\mu(t)\right)} \right| dt \right\|_2, \tag{6}$$

where $\log(\cdot)$ scales the contribution of each pixel along the path according to its semantic relevance, ensuring that regions with higher semantic scores have a more pronounced impact on the trajectory. $S$ is the distance metric denoting the blob score as defined in Eq. (1a). By minimizing $\mathcal{L}_{sc}$, we can avoid high gradient directions that lead to unstable and less interpretable model explanations, ensuring a focus on semantically meaningful regions. The optimal solution path to minimize $\mathcal{L}_{sc}$ is unbounded and can deviate infinitely off the baseline-input region. We therefore incorporate an additional $\ell_2$ norm constraint $\mathcal{L}_{len}$ that minimizes the path length $\mu(t)$ relative to the straight line path $\mu_{\text{str}}(t)$. This proximity reduces the risk of traversing out-of-distribution areas that could lead to unreliable attributions. Moreover, shorter paths inherently promote stability and computational efficiency, ensuring the explanations are both robust and practically feasible. Overall, the objective is to identify the optimal path $\mu^*$:

$$\mu^* = \arg\min_{\mu} \mathcal{L}_{sc} + \lambda \mathcal{L}_{len}, \tag{7}$$

where $\lambda$ is the balancing coefficient, which is converted to an implicit expression by our path design in practice.

Building on the above objectives, we establish specific rules for directing feature transitions. We partition the original path from the baseline to the input into $M$ smaller segments. For each segment, we apply a predefined rule to guide the relevant feature from the beginning to the end of the segment. This strategy ensures that each step of the trajectory of features is controlled and consistent with the path length objectives. In the $m$-th segment, we identify the set of features $\mathbb{U}^{(m)}$, that have not yet achieved their values in the input $x_i$:

$$\mathbb{U}^{(m)} = \left\{ i | \hat{x}_i^{(m)} \neq x_i \right\}, \tag{8}$$

where $\hat{x}_i^{(m)}$ represents the current value of feature $i$ at segment $m$. To refine our selection within this set, we establish criteria $g_i^{(m)}$ for choosing features based on their partial derivatives and semantic scores, and select a subset $\mathbb{P}^{(m)}$:

$$g_i^{(m)} = \log(1 + s_i) \left| \frac{\partial f^d(X)}{\partial x_i} \right|, \tag{9}$$

$$\mathbb{P}^{(m)} = \left\{ i | g_i^{(m)} \leqslant y(\alpha, \{ g_i^{(m)} | i \in \mathbb{U}^{(m)} \}) \right\}, \tag{10}$$

where the specific threshold condition is defined by $y(\cdot)$, with the parameter $\alpha$. For example, by setting $\alpha = 0.05$, the function $y$ can be designed to identify the threshold such that only the features whose $g_i^{(m)}$ values are in the lowest 5% among those in $\mathbb{U}^{(m)}$ are selected to form $\mathbb{P}^{(m)}$. Finally, for the $m$-th segment we define the starting point $\hat{X}_0^{(m)}$, the ending point $\hat{X}_1^{(m)}$, and the directional derivative of the path at those features that satisfy $i \in \mathbb{P}^{(m)}$:

$$\begin{cases} \hat{X}_0^{(m)} = \bar{X} + \dfrac{m-1}{M} \left( X - \bar{X} \right), \\ \hat{X}_1^{(m)} = \bar{X} + \dfrac{m}{M} \left( X - \bar{X} \right), \\ \dfrac{\partial \mu_i(\xi_k)}{\partial k} = \dfrac{1}{M} \Delta x_i \left( -\dfrac{\pi}{K} \sin \left( \pi \dfrac{k + 1/2}{K} \right) \right), \end{cases} \tag{11}$$

where $K$ is the number of sample points used in the Gauss-Legendre quadrature. $k$ means the $k$-th sample point of the Gauss-Legendre quadrature. $M$ is the number of segments between the baseline and the input. In general, the proposed SIAS defines a path that progresses through several steps. At each step, it identifies and selects the subset of features that exhibit the lowest values according to semantic-informed partial derivatives. These selected features are then incrementally moved to more closely match their corresponding values in the original input, while all other features are left unchanged. As the intensity of features align completely with those in the input, they are no longer adjusted. By employing the semantic-informed integration paths, we achieve attribution results that are more relevant to autonomous driving decisions, reduce noise in semantically incoherent regions, and enhance the clarity of attribution explanations.

## V. Experimental Results and Discussions

### A. Implementation Details

For our experiments, we utilize the BDD100k dataset, which comprises 100k images resized to 512 × 256. Additionally, we incorporate the BDD-OIA, an extension of 20k scenes from BDD100k, annotated with binary labels indicating possible actions for the ego-vehicle: "Forward," "Brake," "Turn Left," and "Turn Right." Note that actions such as "Turn Right" and "Forward" are not mutually exclusive in this context. For comparative analysis, we utilize the validation set of the BDD dataset. For experiments requiring semantic segmentation labels, we use the 1k validation subset that includes detailed semantic-level annotations.

We train a multi-label binary DenseNet on BDD-OIA as the autonomous driving model being explained, following the implementation previously described in [59]. The advantage of choosing DenseNet lies in its complex inter-layer connectivity, which enhances the testing of attributional computational modeling capabilities. Because attribution methods that perform well on DNNs with complex structures are considered more robust and have better generalization capabilities. The DenseNet trained on BDD-OIA can be considered as an end-to-end model trained with an imitation learning-style strategy. We train a BlobGAN on the BDD100k dataset, the most training settings remain same as in the original paper [17]. To adapt BlobGAN for generating rectangular images, we modify the feature grid size from 16×16 to 8×16 and empirically increase the number of blob types to 40 to match the object classes of panoptic segmentation labels in the BDD dataset. We set the structural and style feature sizes at 256 to manage complexity effectively. For the stopping criterion, one of the most important hyperparameters for training a BlobGAN, we chose the Fréchet Inception Distance (FID) [60] metric and stopped training when it dropped less than 2% within 5000 steps. The model reached stable convergence after approximately 20 days of training on a server equipped with two NVIDIA RTX A6000 GPUs, using a batch size of 12.

For attribution computation in image space, we set the threshold condition parameter $\alpha = 0.1$, the sample points $K = 30$, and the segment number $M = 10$. This configuration is sufficient for the attributions to approximate sum up to the specific output score for our case, *i.e.*, roughly satisfying the "efficiency" axiom of Aumann-Shapley values as stated in [14]. For blob attribution computation, we maintain the same settings while omitting the blob score term in Eq. (6) to obtain the integration path.

### B. Semantic Attribution Explanations

In this section, we discuss our explanation results and show its effectiveness in identifying defects in the autonomous driving model. Fig. 5 presents our explanations, which include a full attribution map, semantic blob maps aligned with a bar chart detailing the ratios and semantics of blob attributions, and corresponding semantic attribution maps. The semantic attribution maps illustrate the contribution of specific image regions to the driving decision. These maps are generated by intersecting with the full attribution map and the pixel changes upon the removal of respective blobs. Additionally, the blob attributions quantify the impact of respective semantics on the decision-making process.

Attribution scores provide only partial insights into the decision-making process; they indicate where the model is

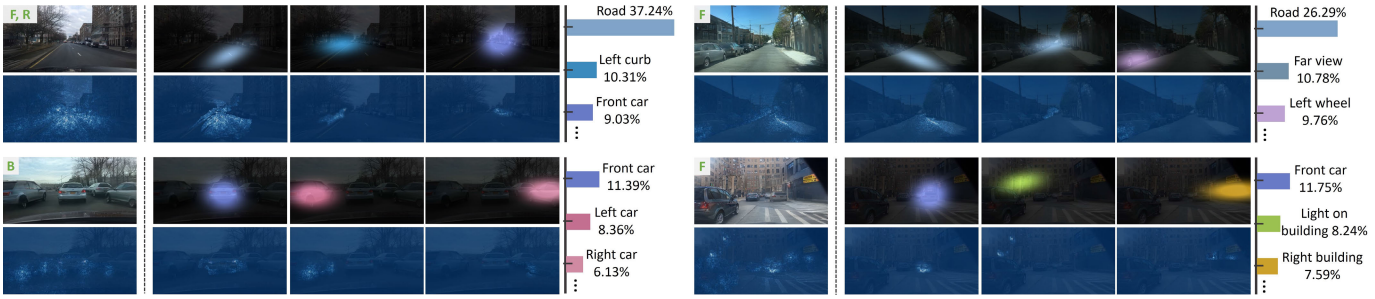



Fig. 5. Semantic attribution explanations. The characters in the original images stand for the driving decisions: "Forward," "Brake," "turn Left," and "turn Right." In each set of explanations, the bottom left displays the full attribution scores. The top right shows the main semantic blobs, which are determined by blob attributions. The semantics of these blobs, along with their attribution ratios, are detailed in the accompanying bar chart. Corresponding semantic attributions, which relate to the upper blob maps, are presented in the bottom right.

focusing without specifying what information is actually used. For instance, in the top left explanation, the full attribution map highlights several key regions relevant to the decision. However, it remains ambiguous whether features such as the road or the curb are decisive in the decision, or to what extent these objects contribute. Moreover, a full attribution map often encompasses multiple objects, which can lead to misinterpretations of the explanation. Consequently, the challenging task of discerning which specific features influence the decision ultimately falls to the human user, potentially leading to false conclusions.

By integrating semantics with attributions, we enhance semantic understanding and bridge the interpretation gap. In the given example, the blob maps and blob attributions reveal that the semantics associated with the "Road," "Left Curb," and "Front Car" play crucial roles in the decision-making process. The corresponding semantic attribution maps help to localize regions for each relevant semantic, providing clearer insights into their contributions. Similarly, in the bottom left explanation, it becomes evident that the decision to "Brake" is due to the presence of cars in three directions. Without semantic-informed attributions, interpreting a clear reason for the model decision to "Brake" would be challenging based solely on traditional attribution scores.

Semantic attributions enable more effective detection of model biases. For instance, in the bottom right explanation of Fig. 5, semantic attributions suggest that a green reflection on a building mistakenly influences the model to decide to move forward. Remarkably, this reflection is not a traffic light, yet the model fails to recognize this, indicating a shortcoming in its ability to recognize contextual nuances. To validate this observation, we introduce red circles in the front view, as shown in Fig. 6. The attribution heatmaps indicate that the red circle regions significantly influence the decision-making process. However, it remains unclear whether these red circles represent the semantics of a traffic light for the autonomous driving model. To clarify this, we generate the semantic blob maps as shown in the bottom row of Fig. 6. The blob in these maps corresponds to the traffic light semantic, confirming that the red circles are mistakenly interpreted as traffic lights by the model, thus leading to incorrect decisions. This modification leads to 27.6% of decisions being incorrect across the validation set. This high rate of mistakes suggests

Fig. 6. Samples leading to erroneous decisions and their attributions. When red circle perturbations are introduced, the model mistakenly decides to initiate a braking action. The attribution maps confirm that these misjudgments are directly influenced by the perturbations, indicating a bias of the model to specific visual stimuli that can lead to incorrect responses.

that the autonomous driving model relies predominantly on color and shape, rather than understanding the actual traffic lights in context. Such a bias could potentially compromise the model performance and pose risks in practical scenarios. By leveraging our semantic attribution explanations, we can identify and understand these biases which could be critical for the development of DNN-based autonomous driving models.

### C. Attribution Semantic Assessment

In this section, we evaluate attribution methods from a semantic analysis perspective to determine whether they effectively capture the semantic information inherent in traffic scenes. Our analysis involves a comparison with seven leading attribution methods: GradShap for autonomous driving [16], IDGI [28], VisualBackProp [27], GuidedGradCAM (an advanced version of GradCAM utilized by [44]), Prop-Shapley [26], AS-QA [14], and LRP-$z^{\mathrm{B}}$ [29]. We explore various computational combinations of LRP using our decision model. Interestingly, we find that LRP-$z^{\mathrm{B}}$, although originally designed for the medical domain, adapts well to autonomous driving models with slight modifications. In particular, our tests show that the default LRP method struggles to produce reasonable attribution results in our driving model. Consequently, we opted for LRP-$z^{\mathrm{B}}$ in our comparison experiments. Given that these attributions methods primarily focus on pixel-level explanations, our comparative experiments are mainly conducted in image space.

We introduce a property comparison to highlight the distinctive attributes of different attribution methods as well

TABLE I
OVERVIEW OF THE RELATED WORKS AND COMPARISON WITH OUR PROPOSAL

| | GradShap | IDGI | VisualBackProp | GuidedGradCAM | PropShapley | AS-QA | LRP-$z^{\mathrm{B}}$ | Ours |
|---|---|---|---|---|---|---|---|---|
| Scenario semantic perception | × | × | × | × | × | × | × | √ |
| Semantic explanation | × | × | × | × | × | △ | × | √ |
| Blob-level attribution | √ | √ | △ | √ | △ | √ | √ | √ |
| No modification on propagation rules | √ | √ | × | √ | × | √ | × | √ |

as the unique advantages of our proposal inspired by [61]. The results are shown in Table I, where a check mark (√) indicates that the property is fully possessed; a triangle (△) suggests that the property is partially possessed; and a cross (×) denotes that the property is not possessed at all. Only our method incorporates semantic perception directly into the generation of explanations, enabling it to produce explanations equipped with semantic understanding. AS-QA can generate implicit semantic explanations through attribution visualization, as discussed in [62]. While most methods can be adapted to perform blob-level attribution calculations when integrated with generative models, specific methods like VisualBackProp and LRP require adjustments to align with the architecture of the generative models. Furthermore, although all these attribution methods rely on backpropagation, methods such as VisualBackProp, PropShapley, and LRP redefine the chain rule, necessitating individual adjustments to the propagation rules. This requirement significantly increases the practical cost and complexity when applying to more sophisticated models.

Evaluating semantic information is challenging due to the difficulty in directly linking attributions to semantics within images. Current attribution explanations primarily emphasize precise causal computations and overlook the integration of semantic information. To the best of our knowledge, there is no existing metric specifically designed to evaluate the semantic information contained in attribution results. To address this, we design two metrics for semantic evaluation: one for scenarios with segmentation labels and another for cases without.

*1) Semantic Assessment for Labeled Data:* First, we use semantic segmentation labels from the BDD dataset to implement the indirect evaluation metric. Initially, we normalize the attribution results of various methods to the output score using the "efficiency" axiom. We then correlate the attribution coordinates with semantic labels to identify the most critical semantic labels. If a majority of the methods highlight a particular label, we consider this semantic element pivotal for decision-making and exclude the corresponding attribution scores within this region. This process is repeated to identify the top 1 to 5 most significant semantic elements using the remaining attribution scores. We then sum the attribution values associated with these semantic regions and divide them by the output score to measure the concentration of attributions in these areas. Table II show the semantic concentration results for varying numbers of semantic elements, ranging from 1 to 5. Larger values signify that the attribution results are more focused within semantic segmentation regions critical to decision-making, as defined by the semantic labels of the BDD dataset. The experimental results show that our method exhibits superior semantic correspondence among all tested methods. Our proposed SIAS is available in three versions, each with different segmentation options: $M = \{5, 10, 20\}$.

These results also reveal that using only one semantic element for evaluation introduces more randomness, disproportionately impacting the performance of certain methods. For instance, GradShap, typically a strong performer, exhibits worse results when considering only one semantic element. Conversely, incorporating three or more semantic elements diminishes the discriminatory power of the evaluation. This is because the selected elements often occupy a substantial portion of an image, leading to high concentration scores for most methods.

Fig. 7 provides two examples that demonstrate the observation outlined above. The leftmost column displays the attribution results, the middle column highlights semantically contributing semantics, and the rightmost column shows less contributing semantics. To determine contributing semantics, we aggregate the attribution results from all methods, considering a semantic contributing only if multiple attribution methods agree. Conversely, less contributing semantics are those detected only by few methods. We can find focusing solely on the most contributing semantics might involve very small image regions. For example, in the top row of Fig. 7, the most contributing semantic element is "traffic light," occupying a very small area within the image. Many methods struggle to precisely localize such small regions, resulting in poor numerical scores. On the other hand, introducing more semantics may overly expand the image coverage, causing the evaluation scores to become too high and lose their discriminatory power. For instance, as shown in the bottom row of Fig. 7, the semantic regions encompassing "car," "road," and "sky," collectively cover almost the entire image. This dominance could potentially undermine the discriminative power of the evaluation metric. Therefore, based on these observations, we designate two semantic elements as the major reference for our evaluation metric. This metric serves as the primary basis for selecting the most effective method for further experiments.

*2) Semantic Assessment for Unlabeled Data:* In instances without available semantic labels, we have developed an extended evaluation metric. We apply mean shift clustering to the coordinates of attribution scores exceeding the 0.1% quantile, identify the centroids, and determine the nearest $B$ blobs. Removing these blobs results in a modified image devoid of the identified semantics. We then assess if this new image alters the model's decision-making and calculate the percentage of decisions that change. A change in the decision indicates that the removed semantics are crucial, demonstrating that the attribution method not only accurately

TABLE II
ATTRIBUTION SEMANTIC CONCENTRATION RATIO

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| GradShap | 17.3 | 46.2 | 70.3 | 89.8 | 94.6 |
| IDGI | 24.2 | 43.7 | 72.1 | 90.4 | 95.1 |
| VisualBackProp | 15.8 | 37.3 | 69.1 | 90.3 | 97.3 |
| GuidedGradCAM | 14.7 | 38.4 | 68.2 | 88.6 | 95.6 |
| PropShapley | 20.6 | 42.7 | 69.7 | 89.2 | 92.4 |
| AS-QA | 23.5 | 42.8 | 70.2 | 88.2 | 96.2 |
| LRP-$z^{\mathrm{B}}$ | 26.4 | 40.1 | 67.5 | 89.7 | 94.7 |
| SIAS (5) | 27.7 | 48.6 | 74.4 | **94.4** | 96.2 |
| SIAS (10) | **28.3** | **49.7** | **75.1** | 93.3 | **97.5** |
| SIAS (20) | 27.8 | 49.2 | 74.9 | 93.2 | 97.2 |

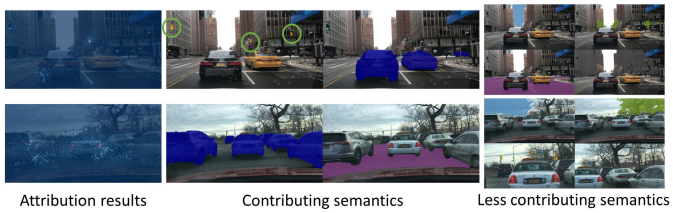



Fig. 7. Samples of contributing and less contributing semantic regions. The attribution values are commonly concentrated on the most important semantic regions. Notably, in the top row, "traffic light" (labeled yellow) is the primary contributing semantic element. This highlights that the most important semantic for decision-making may only occupy a small portion of the overall image. These images are best viewed on screen.

TABLE III
PERCENTAGE OF CHANGES IN DECISION-MAKING

| | B=1 | B=2 | B=3 | B=4 |
|---|---|---|---|---|
| GradShap | 39.4 | 42.7 | 45.4 | 51.9 |
| IDGI | 38.4 | 43.1 | 44.8 | 48.2 |
| VisualBackProp | 31.7 | 33.1 | 36.6 | 39.7 |
| GuidedGradCAM | 32.6 | 35.4 | 38.4 | 40.3 |
| PropShapley | 36.1 | 37.6 | 42.2 | 49.9 |
| AS-QA | 35.7 | 36.4 | 43.1 | 48.5 |
| LRP-$z^{\mathrm{B}}$ | 34.3 | 35.1 | 39.8 | 44.2 |
| SIAS (5) | 42.6 | 46.3 | 50.7 | 53.5 |
| SIAS (10) | 43.3 | **45.8** | **51.2** | **54.6** |
| SIAS (20) | **43.7** | 45.5 | 50.9 | 53.4 |

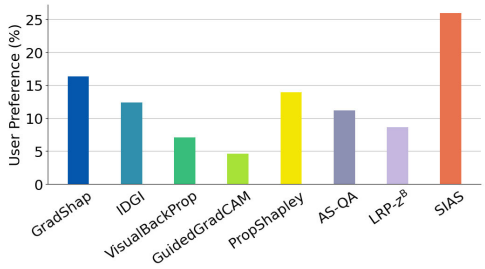



Fig. 8. User preference on attribution results. Higher values suggest that users more clearly understand the connection between attribution results and decision-related semantics, highlighting the effectiveness of the explanations in practical scenarios.

locates semantic information but also exhibits robust semantic expressiveness. Conversely, if the decision remains unchanged, it suggests that the attribution lacks effective semantic expressiveness. These results are detailed in Table III.

Lastly, to validate the practical relevance of these findings, we conducted a user study focusing on the semantic correspondence capabilities of attribution results. We adapted the classical Saliency method [63] to compute blob attributions, pinpointing the most significant semantics. Informed of the decisions influenced by these semantics, we asked participants to assess whether different attribution methods localized effectively to these semantics. Participants from four universities were invited to an online survey, yielding 50 responses. Among these, 19 participants are actively engaged in autonomous driving research, with 9 possessing practical driving experience. The remaining 31 participants, less familiar with autonomous driving, include 11 with driving experience. This diverse pool of respondents helps us assess whether different attribution methods effectively detect crucial semantics in autonomous driving decision-making. Participants were asked to rank the top three attribution methods, assigning scores of 3, 2, and 1, respectively. All other methods received a score of 0. The final scores were then summed to calculate percentage ratios. The user study results, shown in Fig. 8, confirm that our method provides clearer semantic relevance compared to others. This clarity is essential for autonomous driving models, where precise interpretation of critical environmental elements ensures safer decision-making.

TABLE IV
QUANTITATIVE COMPARISON RESULTS BETWEEN OURS AND OTHER ATTRIBUTION METHODS

| | AUC ↑ | AIC ↑ | SIC ↑ | LeRF ↑ | MoRF ↓ |
|---|---|---|---|---|---|
| GradShap | 0.731 | 0.627 | 0.606 | 0.545 | 0.038 |
| IDGI | 0.688 | 0.582 | 0.573 | 0.503 | 0.041 |
| VisualBackProp | 0.610 | 0.524 | 0.539 | 0.497 | 0.045 |
| GuidedGradCAM | 0.606 | 0.518 | 0.521 | 0.486 | 0.048 |
| PropShapley | 0.708 | 0.604 | 0.583 | 0.521 | 0.043 |
| AS-QA | 0.682 | 0.579 | 0.572 | 0.508 | 0.039 |
| LRP-$z^{\mathrm{B}}$ | 0.596 | 0.513 | 0.501 | 0.462 | 0.054 |
| SIAS (5) | 0.754 | 0.641 | 0.625 | 0.587 | 0.026 |
| SIAS (10) | **0.763** | **0.654** | **0.636** | **0.601** | **0.023** |
| SIAS (20) | 0.758 | 0.652 | 0.633 | 0.596 | 0.023 |

### *D. Attribution Result Analysis*

In this section, we analyze the attribution results using several commonly used metrics for comparison. It is important to note that there are currently few attribution methods specifically designed for end-to-end autonomous driving models. Therefore, for our comparative analysis, all methods included are re-implemented on our trained autonomous driving model, adhering to the statements presented in the original papers. Despite this customization, the datasets and comparison metrics used in our experiments are general and publicly accessible, ensuring that our findings remain relevant and broadly applicable across the field.

As shown in Table IV, we quantitatively compare our method with other attribution methods using Area under the ROC Curve (AUC), Accuracy Information Curve (AIC), Softmax Information Curve (SIC) [28], Least-Relevant-Removed-First (LeRF), and Most-Relevant-Removed-First (MoRF) [64]. The AUC is crucial for assessing the ability to discriminate between relevant and irrelevant features, a key factor in safety-critical autonomous systems. A high AUC value


Fig. 9. Comparison of attribution results using heatmaps. Some misleading attribution results are circled in red. In contrast, our attribution results are better localized to key regions and reduce irrelevant noise. These images are best viewed on screen.

indicates that the attribution method is effective at ranking features with a high degree of correspondence to the human-annotated ground truth, meaning that it can reliably distinguish between more and less important features. AIC and SIC provide insights into how well an explanation method can guide the driving model back to high-performance levels (for AIC) and confidence (for SIC) as essential information is incrementally reintroduced. Additionally, LeRF and MoRF evaluate the robustness of attribution methods by observing the impact of sequentially removing the least and most relevant features, respectively. Ideally, autonomous driving models maintaining performance with non-critical features removed (high LeRF values) and showing significant performance drops with crucial features removed (low MoRF values) indicate more accurate and reliable attributions. These metrics collectively assess the overall effectiveness and reliability of explanation methods in complex decision-making scenarios.

The quantitative results in Table IV corroborate the visual observations from Fig. 9. Although visually comparing different attribution results in autonomous driving scenarios is challenging, we could still find some misleading attribution results, even with some methods that generally perform well. For instance, the regions marked by the red circles highlight this issue. Both GradShap and PropShapley generally show good performance; however, they occasionally yield inaccurate results, likely because they solely rely on gradient information and overlook semantic contributions. This focus on gradient information makes them susceptible to problems like vanishing gradients and shattering gradients. In contrast, our proposed method consistently demonstrates strong performance across both evaluation metrics and heatmap visualization results.

## VI. Conclusion

In this work, we apply unsupervised learning to develop a differentiable semantic representation in autonomous driving scenarios, thereby creating a semantically recognizable framework for attribution explanation. We further design the Semantic-Informed Aumann-Shapley (SIAS) method, which integrates semantic blob scores and discrete gradient to determine the path of attribution calculation. This enhances the semantic representational capacity of the attribution computational model and reduces irrelevant noise in the attribution results. Multiple qualitative and quantitative experiments demonstrate that our method effectively facilitates semantic-level explanations and establishes easily understandable connections between decisions in autonomous driving and their attributions.

Despite the advantages mentioned above, our proposal does have some limitations. The semantic blob generation model trained only on the BDD dataset, and its effectiveness on out-of-distribution data remains uncertain. Applying technologies like AI-generated content to achieve more generalized semantic encoding in traffic scenarios is a promising research direction. Additionally, although we have developed evaluation methods tailored to the properties of attribution semantics, there remains a notable gap in general metrics for assessing the semantic information represented by attributions. Further exploration of more universal evaluation methods for attribution analysis is necessary. It is important to acknowledge that while our method effectively addresses attribution challenges in specific autonomous driving models, the emerging field of multi-agent systems presents new complexities and demands [65]. Recent research [66] has begun to explore this area, with a primary focus on counterfactual reasoning as a basis for explanation. However, a key challenge remains: extending attribution methods with robust theoretical guarantees to encompass a broader range of autonomous driving models, especially those characterized by multi-agent interactions. This extension represents a critical frontier in the quest for more comprehensive and reliable explanations in the rapidly evolving landscape of autonomous driving.

## References

[1] Y. Feng, W. Hua, and Y. Sun, "NLE-DM: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 9780–9791, Sep. 2023.

[2] C.-J. Li, Z. Qu, and S.-Y. Wang, "PerspectiveNet: An object detection method with adaptive perspective box network based on density-aware," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 5, pp. 5419–5429, May 2023.

[3] Z. Shen, K. Cai, P. Zhao, and X. Luo, "An interactively motion-assisted network for multiple object tracking in complex traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 1992–2004, Feb. 2024.

[4] T. Huang et al., "Driver lane change intention prediction based on topological graph constructed by driver behaviors and traffic context for human–machine co-driving system," *Transp. Res. C, Emerg. Technol.*, vol. 160, Mar. 2024, Art. no. 104497.

[5] Z. Cao, Y. Gao, J. Bai, Y. Qin, Y. Zheng, and L. Jia, "Efficient dual-stream fusion network for real-time railway scene understanding," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 9442–9452, Aug. 2024.

[6] G. Guo, J. Liu, and X. Sun, "A model decomposition Kalman filter for enhanced localization of land vehicles," *IEEE Trans. Veh. Technol.*, vol. 72, no. 8, pp. 10013–10023, Aug. 2023.

[7] Y. Li, F. Feng, Y. Cai, Z. Li, and M. A. Sotelo, "Localization for intelligent vehicles in underground car parks based on semantic information," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 1317–1332, Feb. 2024.

[8] G. Du, Y. Zou, X. Zhang, Z. Li, and Q. Liu, "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.*, vol. 24, no. 8, pp. 8304–8323, Aug. 2023.

[9] S. Su, X. Ju, C. Xu, and Y. Dai, "Collaborative motion planning based on the improved ant colony algorithm for multiple autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 3, pp. 2792–2802, Mar. 2024.

[10] Y. Wang et al., "Automatic vehicle trajectory data reconstruction at scale," *Transp. Res. C, Emerg. Technol.*, vol. 160, Mar. 2024, Art. no. 104520.

[11] Y. Xue, C. Wang, C. Ding, B. Yu, and S. Cui, "Observer-based event-triggered adaptive platooning control for autonomous vehicles with motion uncertainties," *Transp. Res. C, Emerg. Technol.*, vol. 159, Feb. 2024, Art. no. 104462.

[12] Y. Hu et al., "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 17853–17862.

[13] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. Lopez, "Multimodal end-to-end autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 537–547, Jan. 2022.

[14] R. Shi, T. Li, and Y. Yamaguchi, "Output-targeted baseline for neuron attribution calculation," *Image Vis. Comput.*, vol. 124, Aug. 2022, Art. no. 104516.

[15] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *Int. J. Comput. Vis.*, vol. 130, pp. 2425–2452, Aug. 2022.

[16] M. Li, Y. Wang, H. Sun, Z. Cui, Y. Huang, and H. Chen, "Explaining a machine-learning lane change model with maximum entropy Shapley values," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 6, pp. 3620–3628, Jun. 2023.

[17] D. Epstein, T. Park, R. Zhang, E. Shechtman, and A. A. Efros, "BlobGAN: Spatially disentangled scene representations," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, vol. 13675, 2022, pp. 616–635.

[18] R. J. Aumann and L. S. Shapley, *Values of Non-Atomic Games*. Princeton, NJ, USA: Princeton Univ. Press, 1974.

[19] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," 2018, *arXiv:1805.04687*.

[20] Y. Xu et al., "Explainable object-induced action decision for autonomous vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9520–9529.

[21] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10142–10162, Aug. 2022.

[22] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," 2021, *arXiv:2112.11561*.

[23] M. Sacha, D. Rymarczyk, L. Struski, J. Tabor, and B. Zielinski, "ProtoSeg: Interpretable semantic segmentation with prototypical parts," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1481–1492.

[24] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3145–3153.

[25] A. Oseni et al., "An explainable deep learning framework for resilient intrusion detection in IoT-enabled transportation networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 1000–1014, Jan. 2023.

[26] H. Chen, S. M. Lundberg, and S.-I. Lee, "Explaining a series of models by propagating Shapley values," *Nature Commun.*, vol. 13, no. 1, p. 4512, Aug. 2022.

[27] M. Bojarski et al., "VisualBackProp: Efficient visualization of CNNs for autonomous driving," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1–8.

[28] R. Yang, B. Wang, and M. Bilgic, "IDGI: A framework to eliminate explanation noise from integrated gradients," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23725–23734.

[29] P. R. A. S. Bassi, S. S. J. Dertkigil, and A. Cavalli, "Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization," *Nature Commun.*, vol. 15, no. 1, p. 291, Jan. 2024.

[30] A. Kapishnikov, T. Bolukbasi, F. Viegas, and M. Terry, "XRAI: Better attributions through regions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4947–4956.

[31] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[32] H. Chen, I. C. Covert, S. M. Lundberg, and S.-I. Lee, "Algorithms to estimate Shapley value feature attributions," *Nature Mach. Intell.*, vol. 5, no. 6, pp. 590–601, May 2023.

[33] D. Lundström, T. Huang, and M. Razaviyayn, "A rigorous study of integrated gradients method and extensions to internal neuron attributions," in *Proc. Int. Conf. Mach. Learn.*, vol. 162, 2022, pp. 14485–14508.

[34] J. D. Janizek et al., "Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models," *Nature Biomed. Eng.*, vol. 7, no. 6, pp. 811–829, May 2023.

[35] C. Kim et al., "Transparent medical image AI via an image–text foundation model grounded in medical literature," *Nature Med.*, vol. 30, pp. 1–12, Apr. 2024.

[36] P. S. Chib and P. Singh, "Recent advancements in end-to-end autonomous driving using deep learning: A survey," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 103–118, Jan. 2024.

[37] L. L. Li et al., "End-to-end contextual perception and prediction with interaction transformer," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 5784–5791.

[38] Z. Yuan, X. Song, L. Bai, Z. Wang, and W. Ouyang, "Temporal-channel transformer for 3D LiDAR-based video object detection for autonomous driving," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2068–2078, Apr. 2022.

[39] H. Shenga et al., "Improving 3D object detection with channel-wise transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2723–2732.

[40] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. D. Reid, H. Rezatofighi, and S. Savarese, "Social-BiGAT: Multimodal trajectory forecasting using Bicycle-GAN and graph attention networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 137–146.

[41] B. Ivanovic and M. Pavone, "The Trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2375–2384.

[42] B. Wei, M. Ren, W. Zeng, M. Liang, B. Yang, and R. Urtasun, "Perceive, attend, and drive: Learning spatial attention for safe self-driving," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 4875–4881.

[43] C. Gou, Y. Zhou, and D. Li, "Driver attention prediction based on convolution and transformers," *J. Supercomput.*, vol. 78, no. 6, pp. 8268–8284, Jan. 2022.

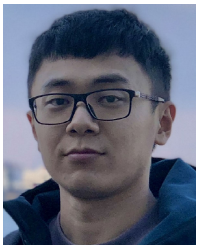
[44] K. Ishihara, A. Kanervisto, J. Miura, and V. Hautamäki, "Multi-task learning with attention for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2902–2911.

[45] Q. Zhang, M. Tang, R. Geng, F. Chen, R. Xin, and L. Wang, "MMFN: Multi-modal-fusion-net for end-to-end driving," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 8638–8643.

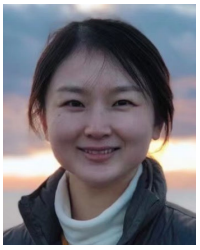
[46] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "TransFuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12878–12895, Nov. 2023.

[47] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7077–7087.

[48] P. Wang and N. Vasconcelos, "A generalized explanation framework for visualization of deep learning model predictions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9265–9283, Aug. 2023.
[49] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, Dec. 2021.
[50] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.
[51] M. Zemni, M. Chen, É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "OCTET: Object-aware counterfactual explanations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15062–15071.
[52] A. Samadi, A. Shirian, K. Koufos, K. Debattista, and M. Dianati, "SAFE: Saliency-aware counterfactual explanations for DNN-based automated driving systems," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 5655–5662.
[53] A. Samadi, K. Koufos, K. Debattista, and M. Dianati, "SAFE-RL: Saliency-aware counterfactual explainer for deep reinforcement learning policies," *IEEE Robot. Autom. Lett.*, vol. 9, no. 11, pp. 9994–10001, Nov. 2024.
[54] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11618–11628.
[55] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7699–7707.
[56] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[57] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape open dataset for autonomous driving and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020.
[58] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, Jan. 2017.
[59] X. Du, Z. Wang, M. Cai, and Y. Li, "VOS: Learning what you don't know by virtual outlier synthesis," 2022, *arXiv:2202.01197*.
[60] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.
[61] G. Basile, D. G. Lui, A. Petrillo, and S. Santini, "Deep deterministic policy gradient virtual coupling control for the coordination and manoeuvring of heterogeneous uncertain nonlinear high-speed trains," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108120.
[62] R. Shi, T. Li, L. Zhang, and Y. Yamaguchi, "Visualization comparison of vision transformers and convolutional neural networks," *IEEE Trans. Multimedia*, vol. 26, pp. 2327–2339, 2024.
[63] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.
[64] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," 2020, *arXiv:2001.00396*.
[65] A. Kuznietsov, B. Gyevnar, C. Wang, S. Peters, and S. V. Albrecht, "Explainable AI for safe and trustworthy autonomous driving: A systematic review," 2024, *arXiv:2402.10086*.
[66] B. Gyevnar, C. Wang, C. G. Lucas, S. B. Cohen, and S. V. Albrecht, "Causal explanations for sequential decision-making in multi-agent systems," in *Proc. Int. Conf. Auton. Agents Multiagent Syst.*, 2024, pp. 771–779.



**Rui Shi** received the Ph.D. degree in graphic and computer sciences from The University of Tokyo, Tokyo, Japan, in 2022. He was a Visiting Researcher with the Department of General Systems Studies, The University of Tokyo. He is currently a Lecturer with the School of Information Science and Technology, Beijing University of Technology, Beijing, China. His current research interests include autonomous driving, neural networks, and explainable artificial intelligence.



**Tianxing Li** received the Ph.D. degree in graphic and computer sciences from The University of Tokyo, Tokyo, Japan, in 2021. She is currently a Lecturer with the College of Computer Science, Beijing University of Technology, Beijing, China. Her current research interests include neural networks and computer graphics.

**Yasushi Yamaguchi** (Member, IEEE) received the Ph.D. degree in information engineering from The University of Tokyo, Tokyo, Japan, in 1988. He is currently a Professor with the Graduate School of Arts and Sciences, The University of Tokyo. His research interests include image processing, computer graphics, and visual illusion, including image editing, computer-aided geometric design, visual cryptography, and hybrid image. He was a former President of the International Society for Geometry and Graphics.

**Liguo Zhang** (Member, IEEE) received the Ph.D. degree in control theory and applications from Beijing University of Technology (BJUT), Beijing, China, in 2006. Since 2014, he has been a Full Professor with the School of Electronic Information and Control Engineering, BJUT. He is currently the Deputy Director of the School of Information Science and Technology, BJUT. His research interests include hybrid systems, intelligent systems, and control of distributed parameter systems. He is an Associate Editor of the *IMA Journal of Mathematical Control and Information* and the Guest Editor of the *International Journal of Distributed Sensor Networks*.