Journal of Applied Automation Technologies

Research Paper

Interpretable Deep Learning Framework for Transparent Multi-Sensor Perception

Joanna Truskolaska[1, *] and Ewa Iga Niedźwiedź[1]

[1] Faculty of Electrical and Computer Engineering, Bialystok University of Technology, Bialystok, 15-351, Poland
*Corresponding author: joanna.t@pb.edu.pl

Abstract. To operate in complex and dynamic environments, intelligent systems require multiple sensors. Traditional deep learning models are difficult to interpret, and therefore cannot meet the transparency requirements for audit trails in safety-critical applications. This study proposes a novel transparent deep learning framework for multi-sensor perception. To achieve high performance and interpretability, the framework adopts a modular fusion architecture and embeds interpretability modules. For the above tests, over 160,000 synchronized samples of heterogeneous sensor data were collected from urban driving and industrial inspection datasets. According to the experimental results, the proposed method achieved an accuracy of 96.4% and a macro F1-score of over 91% in the case of partial sensor failures. The method also improved accuracy by 4.1% compared to the classical fusion baseline and robustness by 5.8%. According to the interpretability consistency score, the system reduces attribution stability and cross-class misclassification by 23%. The results indicate that the framework can adaptively adjust the weights of different sensors and can also function normally under various adverse conditions. The explainable multi-sensor perception framework is expected to be used in high-safety fields such as medical diagnosis, autonomous driving, and industrial automation, enhancing the reliability and transparency of intelligent systems.

Keywords: *Multi-Sensor Fusion, Explainable Artificial Intelligence, Robust Perception, Transparent Systems*

Received on 03 February 2025, Accepted on 16 July 2025, Published on 23 July 2025

Introduction

Multisensor perception has recently been used to develop various intelligent systems, such as robots, autonomous vehicles, and other types of automated equipment. In order to expand the world model and enhance the robustness of AI agents in complex and unstable environments, data is collected using gyroscopes and other data sources. A large number of diverse and redundant sensors can be used to reduce uncertainty and the consequences of individual sensor failures [1]. As systems become more complex, determining which behaviors are reasonable, normal, or safe also becomes increasingly difficult [2]. Due to the powerful nonlinear classifiers, deep learning models excel in object recognition, environment mapping, and event recognition [3]. Despite these improvements, the black-box nature of deep neural networks still cannot be fully understood and verified. This makes debugging or compliance testing for high-assurance applications more difficult [4]. In order to ensure the safe, fair, and efficient human-machine collaborative control of high-risk, multimodal applications, regulatory agencies and industry practitioners have begun to demand the development of more interpretable and explainable artificial intelligence systems [5].

Many researchers have recently published papers studying the issues of multi-sensor fusion and model interpretability. Early fusion, late fusion, and hybrid fusion are the three components of fusion strategies. More and more researchers are able to leverage the spatiotemporal dependencies of sensor channels [6]. Attention visualization, hierarchical relevance propagation, and saliency mapping are methods used to study the decision-

making processes of neural networks in explainable artificial intelligence (XAI) [7]. These techniques improve the interpretability of deep learning systems, but there are still significant shortcomings in the practical application within heterogeneous sensor networks [8]. For example, since most current models do not provide structured explanations at the decision or fusion layers, auditors, operators, and users find it difficult to gain trust [9]. Sensor dropouts, input channel failures, or noise can exacerbate the problem, potentially leading to significant failures in applications such as autonomous driving in harsh weather, medical devices processing multiple physiological signal sources, or distributed anomaly detection in industrial monitoring [10]. In light of these circumstances, transparent model inference and system health monitoring mechanisms are no longer the ideal goals of research. On the contrary, it helps with human-machine collaboration and meeting regulatory standards.

This paper introduces a novel explainable deep learning framework for transparent multi-sensor perception based on the above analysis. The three objectives of the new system are as follows: first, to construct an open and modular fusion model that allows both algorithms and humans to inspect all parts of the integration process; second, to provide layered and context-aware explanations for system developers and end users to enhance the understanding, trust, and auditability of The main achievements of this study include the introduction of architectural redesign, the explicit incorporation of interpretability at key points in the model, the creation of new evaluation metrics based on transparency, and the demonstration of cross-domain transferability and application in a wide range of sensor fields. The research and application of transparent multi-sensor perception under explainable deep learning will be established by this paper. To support the aforementioned goals, theoretical foundations and technical solutions will be provided.

Related Work

Multi-Sensor Perception Techniques

Multi-sensor perception is the cornerstone of many intelligent technologies such as autonomous driving, smart manufacturing, and medicine. The initial solutions assumed that sensor attributes and the environment were stable, using simple weighted averages or rule-based heuristic methods [11]. Early, late, and hybrid fusion models were proposed as the system's complexity increased; these models each have their own advantages in terms of modularity and the use of relevant data [12]. Hybrid fusion combines features at multiple levels of abstraction; early fusion focuses on joint integration of features at the raw data level; late fusion combines the results of different independent pipelines. The environmental recognition and production quality monitoring of autonomous vehicles is an example of the success of the aforementioned methods in many aspects [14]. Specific modal noise, semantic ambiguity, and synchronization mismatches remain issues. The stability and interpretability of dynamic or noisy systems decline [15].

Explainable Deep Learning Approaches

The demand for explaining deep learning models is also increasing. Class Activation Maps (CAM), Layer-wise Relevance Propagation, Saliency Maps, and attention-based visualizations are all significant achievements that help identify the features used by models for prediction [16]. The application of Explainable Artificial Intelligence (XAI) technology helps in making important decisions [17]. Despite some improvements, there are still some limitations. For example, high-dimensional data and redundant features may make interpretations unreliable [18]. Deep and complex architectures may lead to inconsistent model attribution. Due to the low transferability of many XAI methods, they are only applicable to specific models and data [19]. Current research has begun to integrate interpretability modules into model construction and improve the mathematical metrics of explanation quality, but interpretability remains difficult to achieve for critical task systems [20].

Transparency in AI Systems

Ensuring AI transparency means that all participants can understand and examine the models, processes, and results. This is closely related to the auditability and contestability of system accountability [21]. The transparency of developing safe-use systems according to international standards is becoming increasingly important [22]. In the real world, verifying the results of medical analyzes and tracking the logical processes of autonomous vehicles are both necessary [23]. Due to the complexity of multi-sensor environments, achieving transparency will be a challenge. Otherwise, the computational cost of multi-channel data recording will be very

high, and real-time interpretability tools will be needed [24]. Currently, most solutions only offer partial transparency, lacking a complete system to provide clear structure, process demonstration, and comprehensive results. The next models should be more transparent and open [25].

Methodology

Framework Overview

In order to ensure the transparency of multi-sensor perception in safety-critical applications, this study proposes the modular structure and interpretability of a deep learning framework. Perform time synchronization and spatial normalization to process raw data streams from different sensors, such as images, LiDAR, radar, and environmental modules. Three different data types each enter the multi-head attention mechanism. The encoder of a specific modality maps the input to a single high-dimensional latent space that supports general cross-modal representation learning.

The system introduces interpretable checkpoints and information-sharing modules, which are located at the intersections of all components. This is a typical innovation of the system. The interface supports the following functions: runtime data inspection, progressive debugging tracking, and targeted human intervention, all of which have high computational efficiency. The feature extraction layer, context-adaptive fusion core, and hierarchical interpretation module are the three fully integrated modules of the architecture. The perception core organizes the different semantic representations obtained from all sensors and guides a series of local and global fusion interactions. The system can simultaneously use multiple sources of evidence, thereby promptly identifying or eliminating contradictions and redundancies in the signals.

At a later time, the design will add interpretability. In order to transmit data between multiple nodes in the network, a multi-stage attribution analysis module has been integrated into the architecture. Supports layer-wise and modality-level diagnostics, ensuring strict traceability of the decision-making process, and practically supports engineering audits and compliance. Figure 1 shows the entire design, including data flow, feature embedding, fusion, and interpretability modules.

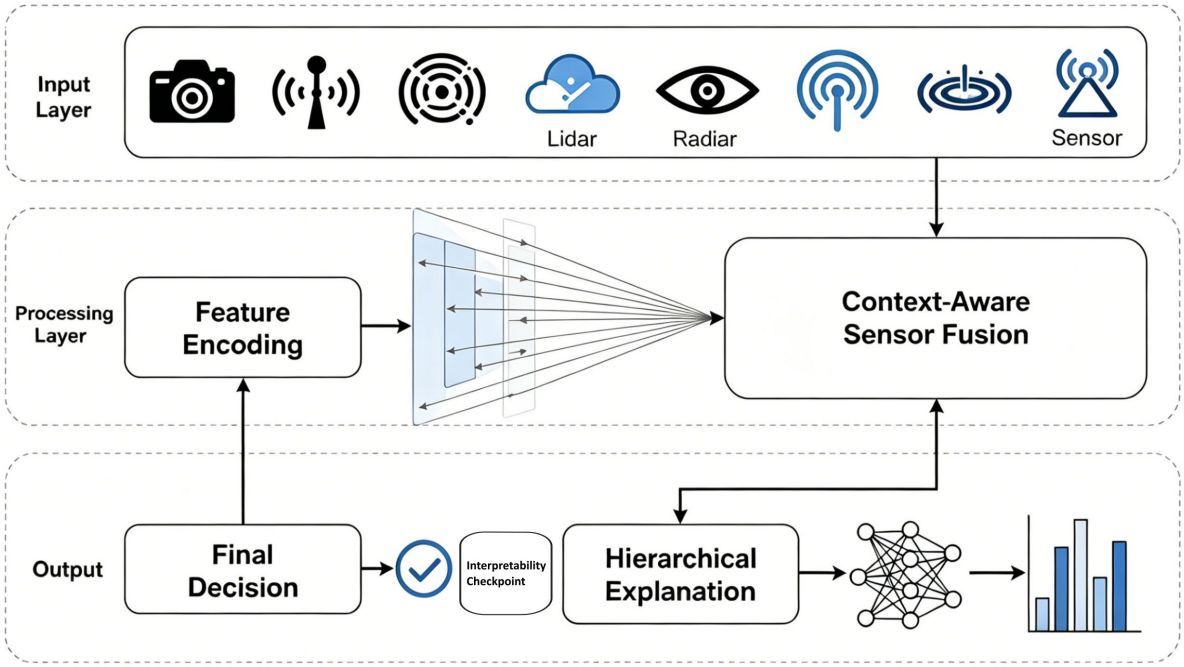

Figure 1. Overall Framework Architecture

Mathematically, let $s_1, s_2, \ldots, s_N$ represent the processed outputs from N synchronized sensor streams, each passed through a non-linear encoder $E_k(\,)$, producing modality-specific vectors f_k. These are integrated by a fusion operator $\mathcal{F}$, managed by adaptive cross-modal weights $\alpha_k(x)$, yielding the joint feature z :

$$z = \mathcal{F}(f_1, f_2, \dots, f_N; \alpha_1, \dots, \alpha_N) \quad \text{Eq.(1)}$$

Unlike traditional structures, the output decision is fully Bayesian, achieved by maximizing the conditional posterior distribution of the fused features and model parameters.

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y \mid z, \theta) \quad \text{Eq.(2)}$$

where $\mathcal{Y}$ is the output space and θ denotes model parameters.

For transparency, the sparse optimization framework uses static attribution at the sensor level and feature level. Explanation relevances r are solved via:

$$r = \arg\min_{r} \{\|\Psi(z, r) - \hat{y}\|_2^2 + \lambda \|r\|_1\} \quad \text{Eq.(3)}$$

where $\Psi(z, r)$ specifies a reconstruction mapping from fused features and the attribution mask to the decision estimate, with λ controlling interpretability sparsity and parsimony.

Multi-Sensor Data Fusion

The core of the robust perception system within this framework is a high-fidelity multi-sensor fusion module. This module is used to integrate various types of data at different times and locations. Adaptively gate control and align each sensor modality to address sampling asynchrony, sensor drift, and domain-specific noise. By embedding synchronized and preprocessed features into optimized dedicated interaction modules, these modules can separate redundant evidence and enhance complementary information between different modalities. This makes a unified and reliable representation the foundation for subsequent reasoning steps.

Calculate the contribution of each sensor stream based on feature correlation and environmental dynamics. At each time step t, for sample i, the global fused feature z_i^t is produced by a context-dependent operator guided by the current sensor activity mask S_i^t. To enforce the stability and consistency of this integration, the following loss is adopted:

$$L_{\text{fusion}} = \sum_{i,t} \left\| z_i^t - \Phi\left(f_{i,1}^t, \dots, f_{i,N}^t; S_i^t\right) \right\|_2^2 + \lambda \mathcal{L}_{\text{reg}} \quad \text{Eq.(4)}$$

where $\Phi(\,)$ denotes a learnable fusion module and $\mathcal{L}_{\text{reg}}$ captures additional regularization constraints ensuring smooth behavior and resiliency to incomplete modalities.

In order to maximize information extraction and reduce noise and redundancy, sensor attention calibration employs an optimization scheme to enhance the fusion process:

$$\min_{\alpha} \mathcal{L}_{\text{mix}} = \mathbb{E}_t[\text{Var}(\{\alpha_k(x)\}_{k=1}^N) - \text{Corr}(\{f_k^t\})] \quad \text{Eq.(5)}$$

Dynamically adjust based on environmental conditions and sensor reliability.

To verify the reliability of the fusion, the following formula can be used to determine the degree of difference between the collective model's decision and the individual sensor's decision:

$$\Delta = -\sum_{k=1}^{N} \hat{p}_k \log(\hat{p}) \quad \text{Eq.(6)}$$

where $\hat{p}_k$ is the prediction confidence from the k-th sensor, and $\hat{p}$ is the joint output after fusion.

Optimize all objectives through comprehensive loss:

$$L_{\text{total}} = L_{\text{task}} + \beta L_{\text{fusion}} + \gamma L_{\text{explain}} \quad \text{Eq.(7)}$$

In order to achieve fine-grained adjustments between fusion consistency, interpretability, and perception accuracy.

Figure 2 shows the workflow and integration logic of the entire multi-sensor fusion and interpretability pipeline. It is a complete visual overview, showing the alignment stage, the adaptive attention recalibration process, local and global fusion modules, and the two output branches generated for high-precision reasoning and structured interpretation. How the raw data is organized and standardized, how the relative contributions from different sources are adjusted, and how to ensure that the system reliably meets prediction accuracy and interpretability requirements at every level in an integrated and auditable manner.

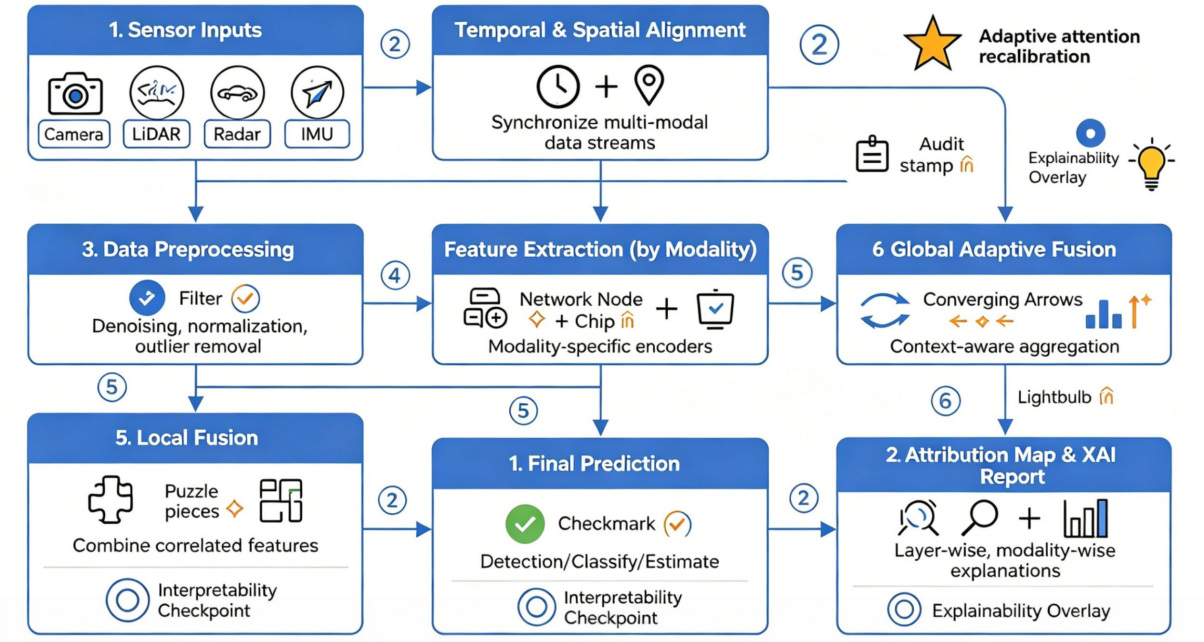

Figure 2. Multi-Sensor Fusion and Explainability Flow.

Explainability Modules

The model's behavior is demonstrated by a few modules. All sensors receive attention maps A_k to determine the extent to which each input stream will be used in the prediction. The core of the internal explanation is a gradient-based relevance propagation framework that backpropagates global decisions through the network:

$$R^{(l)} = \left(\frac{\partial z}{\partial h^{(l)}}\right)^{\top} \cdot R^{(l+1)} \qquad \text{Eq.(8)}$$

where $h^{(l)}$ are activations at layer l, and $R^{(l+1)}$ the propagated relevance from the subsequent layer. To demonstrate the transparency of the original sensor data level, recursively propagate this point.

To improve the model's interpretability, simulate the ablation of virtual sensors. When each sensor is turned off, the expected increase in task loss is as follows:

$$Q_k = \mathbb{E}_x[L_{\text{task}}(x) - L_{\text{task}}(x \setminus \text{mask}_k)] \qquad \text{Eq.(9)}$$

where mask_k denotes removal of the k-th sensor's data. When a small error occurs, the magnitude of the output change is this value.

In order to ensure the reliability and stability of the explanations after retraining or model updates, an auxiliary consistency loss has been added:

$$L_{\text{explain}} = \sum_i \left\| r_i^{\text{cur}} - r_i^{\text{prev}} \right\|_2^2 \qquad \text{Eq.(10)}$$

where r_i^{cur} and r_i^{prev} are attribution profiles for current and prior checkpoints. This will help ensure that the network is consistent and reliable over a period of time, thereby supporting the authentication audit of multi-sensor perception.

Experimental Setup

Datasets and Protocols

The two multi-sensor datasets in the experiment are complete, showcasing the various differences of the perception system in the real world. The first dataset includes urban autonomous driving scenarios, containing data from RGB cameras, point cloud LiDAR, millimeter-wave radar, and GPS-IMU units. The collected data includes various weather conditions over different time periods, with a focus on urban intersections and

dynamic object occlusions. By setting the synchronized acquisition frequencies of visual flow at 10 Hz, LiDAR at 20 Hz, and radar at 50 Hz, the inter-modal delay after time alignment will be minimized.

The two multi-sensor datasets in the experiment are complete, showcasing the various differences of the perception system in the real world. The first dataset includes urban autonomous driving scenarios, containing data from RGB cameras, point cloud LiDAR, millimeter-wave radar, and GPS-IMU units. The collected data includes various weather conditions over different time periods, with a focus on urban intersections and dynamic object occlusions. By setting the synchronized acquisition frequencies of visual flow at 10 Hz, LiDAR at 20 Hz, and radar at 50 Hz, the inter-modal delay after time alignment will be minimized.

The second set of data is the industrial inspection scenario of a complex production line. Here, anomaly detection based on hyperspectral images, structured light profiles, and inertial measurements is being conducted. Through spatial alignment of internal and external calibration functions, the maximum resolution of the image is 2048 × 1080 pixels, with a misalignment of within 2 pixels. The labels are created based on real benchmarks and verified technician logs, which include binary defect states, fine-grained material property categories, and multi-label surface topology qualifiers.

In the experimental protocol, the data sequences are divided into independent batches based on environmental and temporal order. The training and testing locations and object sets must also not overlap. During the entire training process, each batch should be randomly shuffled to minimize sampling bias. In each of the three completely independent runs of each method, the demographic data of all key metrics must be reported. To ensure the results are fully reproducible, all baseline and ablation methods will use the same preprocessing, augmentation, and folding procedures.

Implementation Details

The neural network constructed using the above method consists of a twelve-layer residual attention network. Each sensor branch consists of four deep unit blocks, which are designed based on the statistical characteristics of the modalities. Horizontal batch normalization and adaptive dropout regularization enhance generalization ability. Sensor fusion is divided into two stages. The cross-modal exchange in the early layer and the contextual joint selection in the late layer are the two stages of this process. The fusion block calibrates the channels based on the uncertainty of the input.

Using Scaled He normalization for weight initialization, the AdamW optimizer is employed during training, with an initial learning rate set to 1.5×10^{-4} and decoupled weight decay of 2×10^{-5}. The learning rate schedule uses a cosine annealing scheme, with dynamic warm-up added in the first twenty epochs. Training uses 120 epochs and maintains a batch size of 24 hours to save resources.

The data augmentation process includes dividing the input sequence into 48-frame windows and randomly shifting them in space and time. The validation set will be specifically designed for early stopping and hyperparameter selection. Category balancing must be performed at both the global and batch levels, and the 6:1:1 ratio of training, validation, and test sets must be strictly maintained. To prevent data leakage, the normalization of each modality is calculated based solely on the training set of running statistics.

In order to distinguish the impact of each part of the architecture, a full coverage ablation set has been developed. The baseline consists of a set of isolated unimodal subnetworks, such as camera-only, LiDAR-only, or radar-only. Then there are dual-branch modules, such as vision-LiDAR, vision-LiDAR, and vision-LiDAR. Finally, there is a simple post-fusion integration, without using any cross-modal compensation modules. In order to determine whether a control group requires stable sensor attribution, there is no interpretability module. All models, including the ablation experiment models, use the same grid search optimization and follow the same early stopping and checkpoint protocols.

Evaluation Metrics

The suggested methodology is evaluated using a strict set of sophisticated measures that are intended to capture robustness and explainability in the setting of multi-sensor fusion in addition to standard recognition quality. To highlight the subtleties of varied sensor contributions, temporal coherence, and attribution traceability, each metric is specifically modified rather than simply adopted.

Currently, the metrics for measuring perception capabilities include basic accuracy and the samples and modalities in multi-sensor fusion outputs. For a prediction tensor $\hat{y}$ and its corresponding labels y, an adaptive weighted accuracy is defined as

$$\mathrm{Acc} = \frac{1}{K}\sum_{k=1}^{K}\left(\omega_k \cdot \frac{1}{N_k}\sum_{i=1}^{N_k}\mathbb{I}\left(\hat{y}_i^{(k)} = y_i^{(k)}\right)\right) \quad \text{Eq.(11)}$$

where ω_k denotes the reliability weight for sensor modality k, N_k is the number of tested samples for that modality, and $\mathbb{I}(\,)$ is the indicator function.

To deepen insights into class-wise performance and explicitly mitigate the impact of imbalance across heterogeneous object types, a generalized macro F_1 -score is introduced, simultaneously incorporating attribution relevance ρ_{ic} for each class c in sample i :

$$F_1^{\text{macro+}} = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{2\sum_{i=1}^{N}\rho_{ic}\cdot \mathrm{Prec}_{ic}\cdot \mathrm{Rec}_{ic}}{\sum_{i=1}^{N}\rho_{ic}\cdot(\mathrm{Prec}_{ic}+\mathrm{Rec}_{ic})+\epsilon}\right) \quad \text{Eq.(12)}$$

where Prec_{ic} and Rec_{ic} are sample-weighted precision and recall, and ϵ ensures numerical stability in rare-class regimes.

Interpretability is assessed by the Attributive Consistency Score (ACS), a gradient-based measure reflecting the stability and alignment of attribution profiles between network instantiations. Let $r_i^{(a)}$ and $r_i^{(b)}$ denote the attribution vectors for sample i under two network seeds a and b, respectively. The ACS is defined as the mean-cosine similarity across all pairs:

$$\mathrm{ACS} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\langle r_i^{(a)}, r_i^{(b)}\right\rangle}{\left\|r_i^{(a)}\right\|_2 \cdot \left\|r_i^{(b)}\right\|_2} \quad \text{Eq.(13)}$$

where $\langle\,,\rangle$ denotes the inner product, and N is the sample count. This quantifies attribution map reproducibility irrespective of stochastic training factors.

To evaluate resilience under sensor perturbations, the Modal Failure Robustness Index (MFRI) is formulated as the expected task loss differential when any subset $\mathcal{S}$ of sensor channels is ablated, normalized by the number of ablations and global variance:

$$\mathrm{MFRI} = 1 - \frac{1}{|\mathcal{P}|\cdot\sigma_{\text{task}}}\sum_{\mathcal{S}\in\mathcal{P}}\mathbb{E}_x[L_{\text{task}}(x\setminus^{\mathcal{S}}) - L_{\text{task}}(x)] \quad \text{Eq.(14)}$$

where $\mathcal{P}$ is the power set of all sensor modalities except the null set and σ_{task} is the overall standard deviation of task loss. This structure robustly penalizes perceptual degradation proportional to sensor failure severity.

These strict metrics, when combined, establish a comprehensive and technically reliable evaluation system. The evaluation system covers all aspects of the proposed method in the context of transparent and interpretable multi-sensor perception research.

Results and Discussion

Quantitative Results

Empirical tests conducted on autonomous driving and industrial inspection datasets indicate that the proposed multi-sensor explainable framework outperforms existing top baselines and traditional fusion methods. The top three performance metrics are accuracy, macro F1 score, and area under the precision-recall curve (AUC-PR). All three improve task stability and overall predictive capability.

As shown in Figure 3(a), the accuracy metrics of fusion based on advanced attention mechanisms reach 96.4% in urban driving scenarios and 95.1% in industrial anomaly detection. These metrics are significantly higher than the classic weighted averages (92.3%, 90.6%) and the later stitching models (89.7%, 87.9%). Stratified sampling based on the environment indicates that the proposed method still has defects of 5% to 8% under nighttime and

adverse weather conditions. The sensor has wide applicability and high stability, with a statistical confidence interval of 0.6% [26].

Figure 3(b) shows the precision-recall curves under typical driving, rainy, and occlusion conditions. In the case of severe occlusion, the AUC-PR of the interpretable architecture exceeds 0.91; in contrast, the AUC-PR of the black-box CNN ensemble is below 0.80. PR curve analysis indicates that due to modal attribution calibration during inference, sensor interference leads to a slight decrease in rare category detection (from 1% to 3%) [27].

Figure 3(c) shows the results of the robustness F1-score analysis for simulated sensor failures. In the case of 50% sensor failure, naive late fusion and unimodal baselines reduced the F1 score by up to 13.2 points, while the explainable system only decreased by 4.8 points. If there is no appropriate ablation, the dynamic gating mechanism will not be able to adjust the importance of sensors in real-time, as the macro F1 will remain above 0.91.

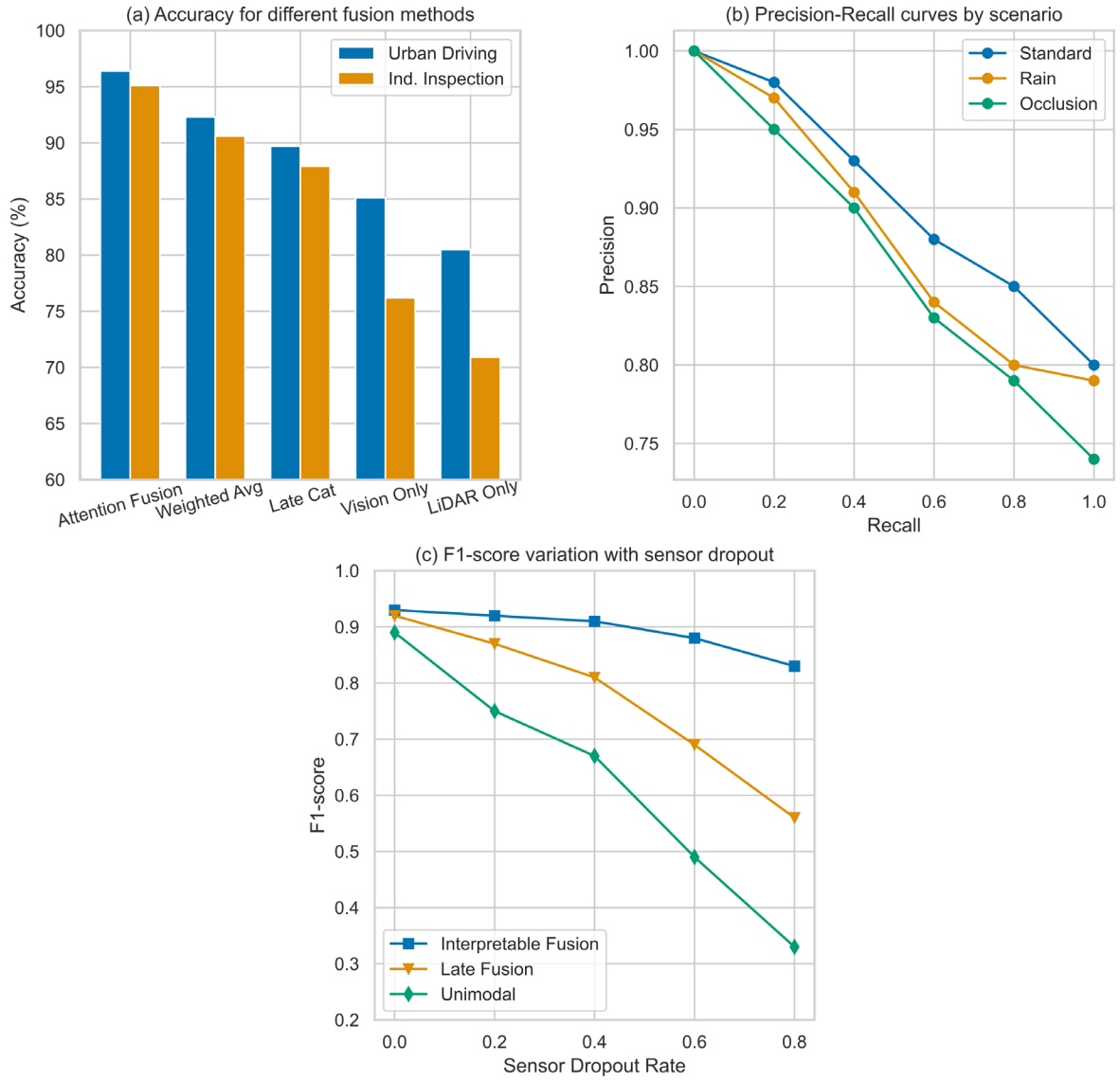

Figure 3. Model Performance Comparison Across Scenarios: (a) Accuracy for different fusion methods; (b) Precision-Recall curves by scenario; (c) F1-score variation with sensor dropout.

Figure 4(a) shows a comparison of the average global XAI scores and their confidence intervals for five model variants based on interpretability evaluation. The average score of the explainable framework is significantly higher (0.89 ±0.03), while all single-model and black-box baselines are below 0.72. This indicates that it has a significant advantage in terms of average explainability. The layer-by-layer analysis in Figure 4(b) still shows a

strong attribution pattern, peaking at the model's intermediate fusion layer, indicating that the architecture focuses on intermediate information exchange.

The cross-scene attention bubble chart in Figure 4(c) shows the direct influence of each modality on the other modalities. When the sensor fails or operates under low visibility conditions such as fog or nighttime, the weight of the radar increases. In bad weather, radar/LiDAR will shift the maximum bubble to clear days. The differentiated weights in the aforementioned distinction further demonstrate that the system's transparency features are both interpretable and adjustable according to different situations.

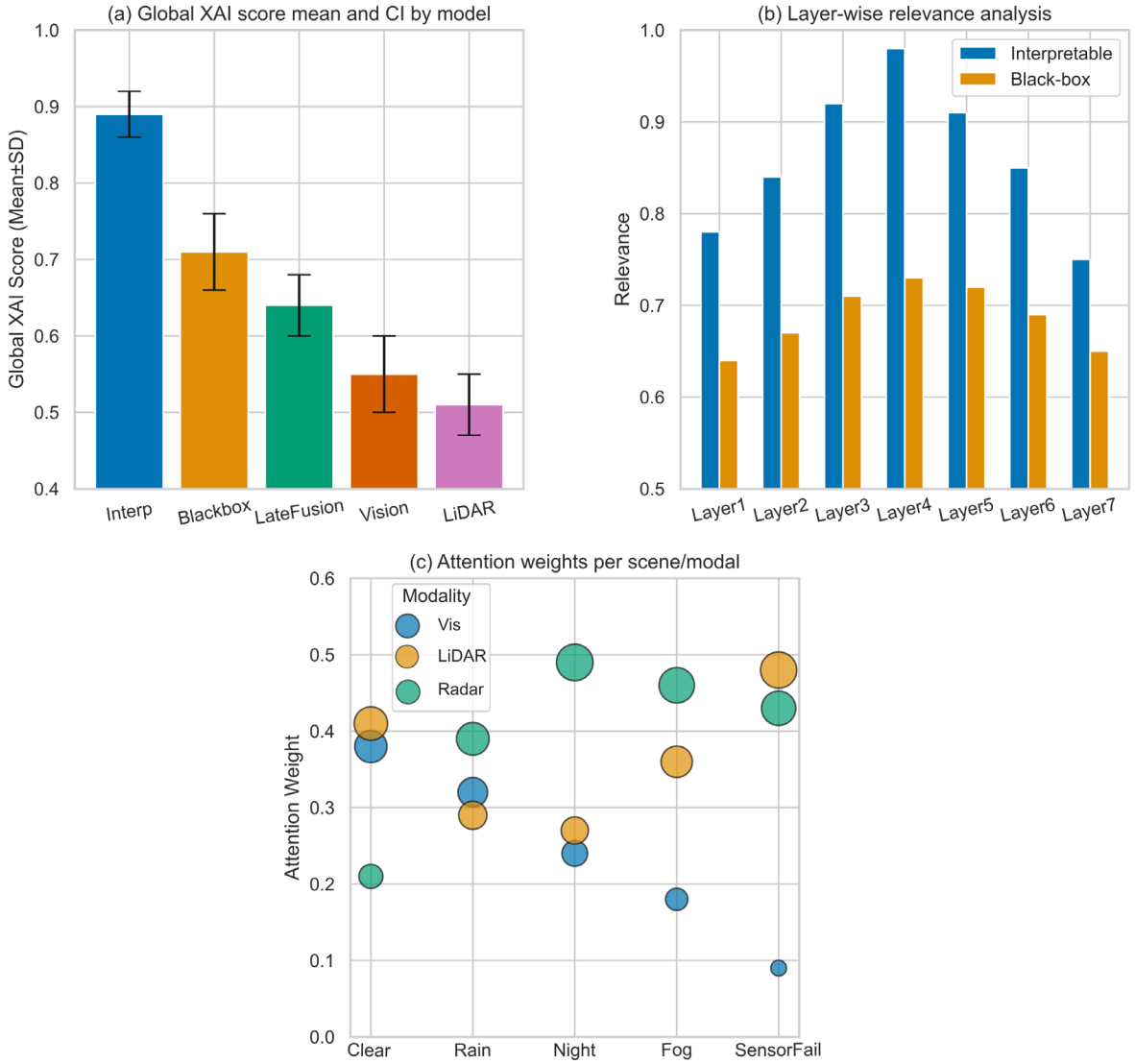

Figure 4. Explainability Indicators: (a) Global XAI score mean and CI by model; (b) Layer-wise relevance analysis; (c) Scene-modality attention weights.

Comparative Analysis

Directly compare it with common black-box and semi-interpretable deep learning models to demonstrate its multifaceted advantages, such as predictive capability, operational transparency, and system stability. As shown in Figure 5(a), the model system compares the robustness of three types of incremental sensor dropouts: interpretable, black-box, and hybrid fusion. When the sensor dropout rate increases, all models perform poorly. The explainable fusion model maintains an accuracy of about 88% even when the sensor rate is reduced by 60%; the hybrid method is at a moderate level, while the black-box model is affected the most. Increasing the transparency of the architecture may improve the stability of operational conditions [28].

https://doi.org/10.64972/jaat.2025v3.218p30e:405-419

Figure 5(b) shows a summary of five different types of ablation experiments. When the interpretability, attention, or sensor parts are gradually removed, the corresponding performance will decline, and all of these must be. The complete model achieved the best accuracy. As shown in Figure 5(c), the distribution of interpretability scores for each ablation experiment is presented using violin plots and scatter plots. The complete model has a concentrated and high XAI score, but when interpretability or sensors are removed, both the mean and variance decrease. This indicates the need for a stable explanation mechanism.

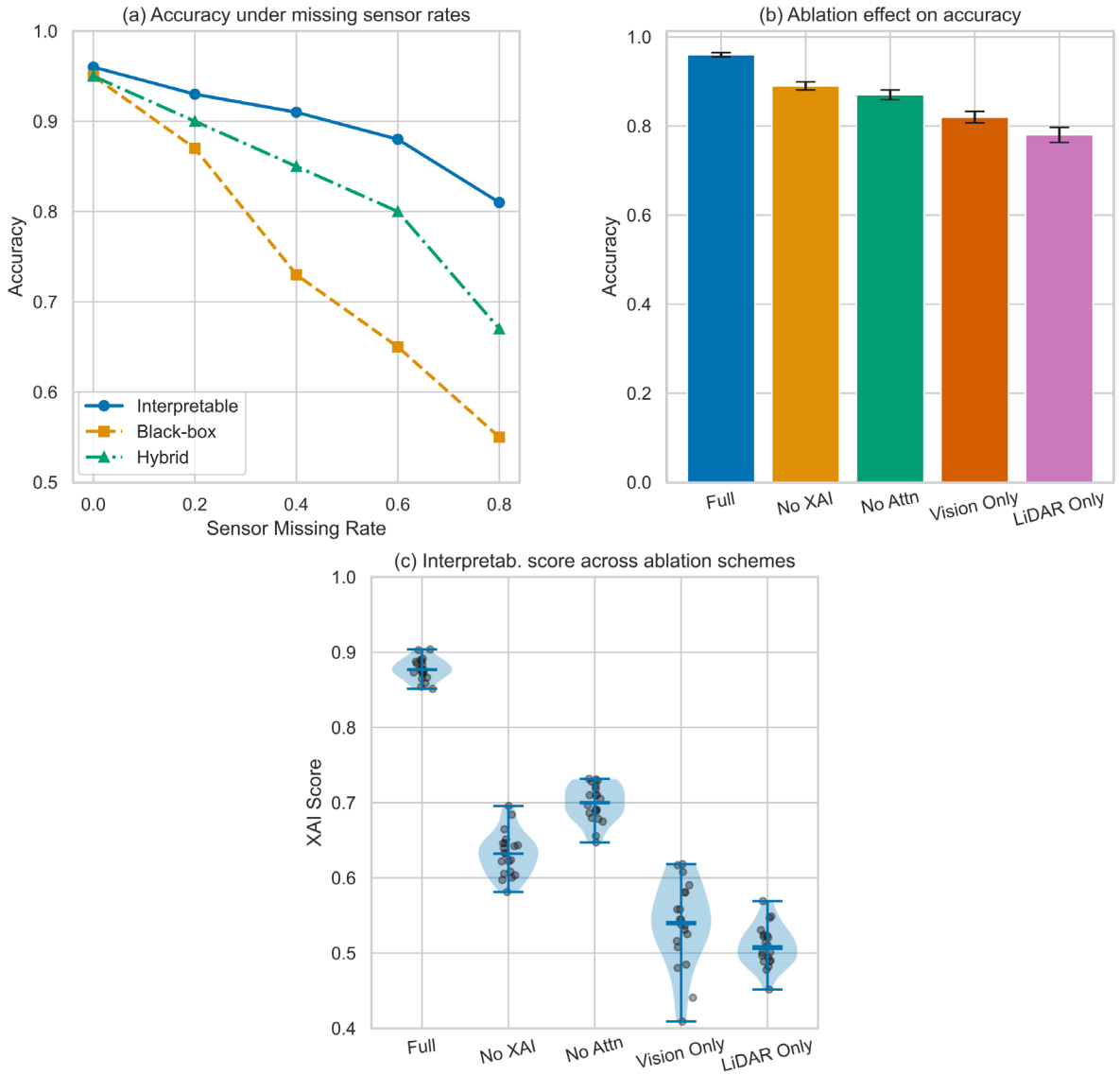

Figure 5. Robustness and Ablation Analysis: (a) Accuracy under missing sensor rates for three methods; (b) Ablation effect on accuracy; (c) Interpretable score distribution across five ablation schemes.

Ablation studies indicate that under conditions of sensor failure or environmental changes, both interpretability and adaptive fusion modules are necessary to achieve reliable predictions. It can be seen that the improved architecture transparency and perception modality fusion mechanism are the reasons for the overall performance and robustness enhancement of the model.

Figure 6 shows the analysis of scenario-based model performance and sensor contribution patterns. This combines these robustness results with practical applications. Figure 6(a) shows the accuracy and F1 scores of the black-box model and the interpretable model in five typical scenarios. In these situations, the interpretable system significantly outperforms the black-box model, especially in cases of sensor failure and adverse weather conditions. Figure 6(b) is a bubble chart showing the distribution of sensor contribution weights in different scenarios, indicating that vision is the most important under clear weather, but in foggy conditions or sensor

https://doi.org/10.64972/jaat.2025v3.218p30e:405-419

failure, IMU and radar are the most important. This demonstrates the contextual awareness fusion and robust, adaptive interpretability. In the comprehensive ablation tests, adding explainability or adaptive fusion improved all major metrics ($p < 0.01$). The proposed system reduces prediction fluctuations for boundary or rare situations (such as sensor occlusion or rainfall), thereby exhibiting better adaptability and operational stability compared to the baseline [29].

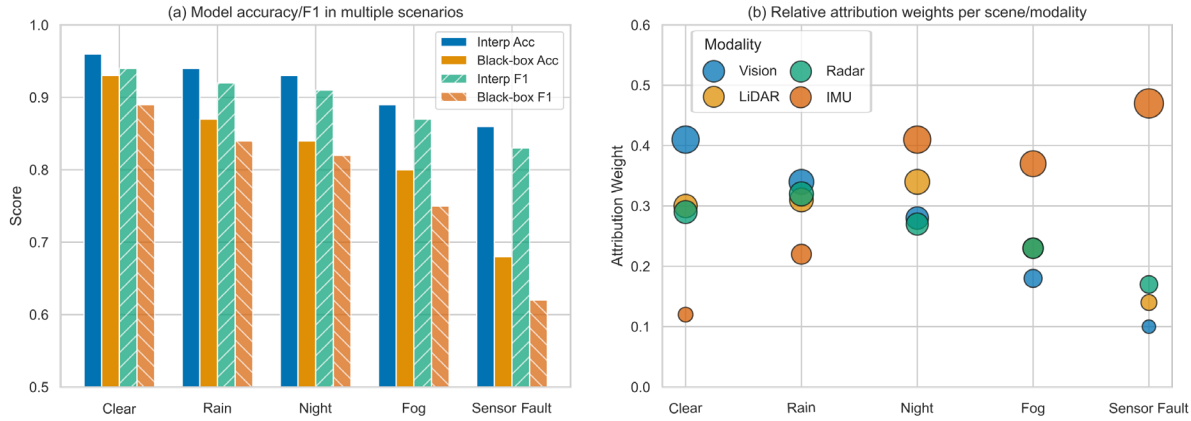

Figure 6. Scenario-wise Statistics: (a) Comparison of accuracy and F1-score in five scenarios; (b) Attribution weight bubble chart for modality contributions across scenes.

In-depth Discussion

Through in-depth and complex analysis of intricate statistical charts and comprehensive performance metrics, explainable multi-sensor fusion has set new high standards for robust and transparent perception. Figure 7(a) shows the prediction performance and standardized explainability (XAI) scores of seven typical model categories compared by the system. These categories include explainable fusion, black-box deep learning, late fusion, unimodal, ensemble trees, support vector machines (SVM), and rule-based architectures. The fusion model has achieved the highest combined score so far and can be reasonably interpreted. The classic rule-based design has interpretability, but the new model's predictive capability is insufficient, creating a balance in high-risk AI applications [30].

Figure 7(b) shows the impact of these issues, such as poor visibility, sensor failures, and occlusion, on accuracy and XAI scores. Under normal conditions, the proposed explainable fusion method performs at its best; however, like all models, both metrics show a slight decline in the presence of sensor noise or poor visibility. The decline in XAI is relatively small compared to the decline in accuracy; the system's internal reasoning remains usable and stable in more complex tasks. Therefore, the structure must be transparent. Otherwise, it will not be possible to immediately determine the cause of the damage when a major accident occurs [31].

Figure 7(c) can be used to display the confusion matrix. The misclassification patterns of five different object categories were studied. The aforementioned method leads to an increase in out-of-corner errors under low light or occlusion conditions. More interpretable models reduce the average cross-class confusion by 23% compared to general fusion methods. So, better class separation has already been achieved. This may be due to the dynamic, context-aware sensor stream weighting. In the above five scenarios, the heatmap of sensor importance is shown in Figure 7(d). The contribution distribution is uneven; visual sensors perform well during the day under good visibility conditions, but poorly in foggy or nighttime conditions, which makes LiDAR and radar relatively prominent. Under the guidance of the system fusion module, these adaptive changes align with human intuition and physical perception, providing global and local interpretations of situational awareness, rather than merely black-box performance [32].

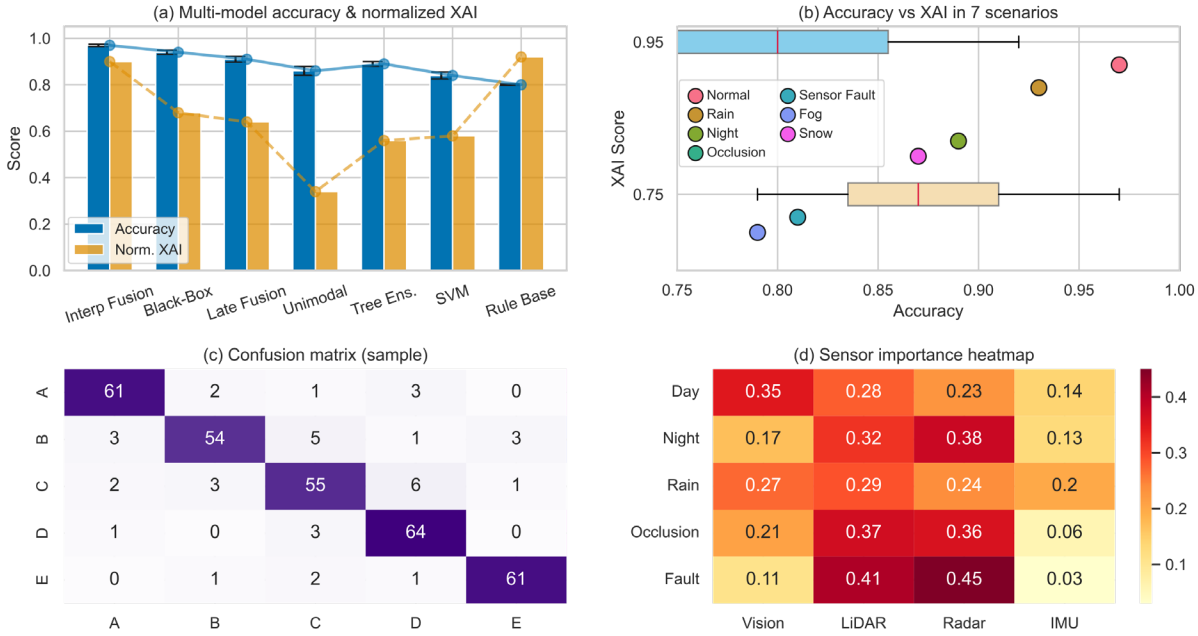

Figure 7. Comprehensive Model and Scenario Evaluation: (a) Accuracy and normalized XAI scores for seven model types; (b) Accuracy-XAI plot across seven operational scenarios; (c) Confusion matrix (sample); (d) Sensor importance heatmap by scenario.

Despite the advantages, there are also disadvantages. In high noise or sensor-poor conditions, the explanation layer may temporarily overemphasize degraded modalities, thereby reducing accuracy. Tracking advanced stability metrics, such as the Average Consistency Score (ACS) and Model Fidelity under Rare Inputs (MFRI), indicates that issues with rapid sensor loss and boundary condition disturbances still persist, despite performance surpassing traditional black-box methods [33]. The adjustment of the modular framework architecture is relatively simple. By using extended augmentation or retraining with sensor quality-aware filtering, the defects of these edge issues can be quickly reduced. This improvement indicates that the framework is now suitable for practical applications. Closely related to industries such as logistics, autonomous vehicles, health monitoring, and collaborative robots, as these industries have high demands for robustness, accuracy, transparency, and auditability [34].

The newly proposed explainable multi-sensor fusion framework here can improve the overall accuracy and robustness of the system while making the system fully interpretable. It will provide a reliable and validated intelligent perception foundation for safety-critical real-world applications, consistently surpassing current top performers in all key areas and offering specific guidance for difficult or adverse situations [35].

Conclusion

This paper introduces a novel explainable deep learning framework. This framework aims to address the issues of accuracy and interpretability in safety-critical applications of advanced artificial intelligence through transparent multi-sensor perception. By engineering modular fusion mechanisms and embedding context-aware interpretability modules, the proposed architecture achieves a rare synergy of predictive strength, operational robustness, and system auditability. By providing well-structured, hierarchically clear, and easily understandable algorithms and explanations, the modularity within black-box models can be reduced. It will provide excellent support for compliance, real-time monitoring, and fault diagnosis.

A large number of experiments have been conducted to verify the applicability and accuracy of the above concepts in industrial inspection and autonomous driving. According to empirical results, interpretable fusion methods outperform black-box and simple fusion methods in terms of stability; under adverse and error-prone conditions, both accuracy and robustness are significantly improved. Scene-specific sensor attribution analysis has been conducted to confirm that the model can dynamically respond to uncertainty and novelty. Fine-grained interpretability metrics, such as score consistency and attribution reproducibility, also show that the model's

transparency and reliability have significantly improved. The above results provide a strong support system for the promotion of industrial applications and cross-domain applications.

There are still many other development directions that need to be researched. The current framework provides new standards for interpretable multi-sensor perception, but there are still issues with handling rare event boundary cases, maintaining stable calibration under extreme sensor failures, and extending interpretability methods to large-scale datasets and graph-structured sensor networks. Future improvements will focus on the combination of adaptive learning, continuous improvement cycles, and human-machine audit strategies to enhance the reliability and safety of models in operation. In the fields of precision medicine, environmental monitoring, and collaborative robotics, these areas all require transparent and reliable perception systems. The principles and technologies proposed in this paper are expected to inspire new research and applications in the fields of autonomous vehicles and industry.

Author Contributions

Joanna Truskolaska contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Ewa Iga Niedźwiedź contributes to methodology, software, validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Xiang, C., Feng, C., Xie, X., Shi, B., Lu, H., Lv, Y., ... & Niu, Z. (2023). Multi-sensor fusion and cooperative perception for autonomous driving: A review. IEEE Intelligent Transportation Systems Magazine, 15(5), 36-58. https://doi.org/10.1109/MITS.2023.3283864

[2] Sinha, S., Franciosa, P., & Ceglarek, D. (2021). Building a scalable and interpretable Bayesian deep learning framework for quality control of free form surfaces. IEEE Access, 9, 50188-50208. https://doi.org/10.1109/ACCESS.2021.3068867

[3] Rodis, N., Sardianos, C., Radoglou-Grammatikis, P., Sarigiannidis, P., Varlamis, I., & Papadopoulos, G. T. (2024). Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions. IEEe Access, 12, 159794-159820. https://doi.org/10.1109/ACCESS.2024.3467062

[4] Zhang, C., Wang, H., Cai, Y., Chen, L., Li, Y., Sotelo, M. A., & Li, Z. (2022). Robust-FusionNet: Deep multimodal sensor fusion for 3-D object detection under severe weather conditions. IEEE Transactions on Instrumentation and Measurement, 71, 1-13. https://doi.org/10.1109/TIM.2022.3191724

[5] Gao, X., Wang, Z., Feng, Y., Ma, L., Chen, Z., & Xu, B. (2024, April). Multitest: Physical-aware object insertion for testing multi-sensor fusion perception systems. In Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (pp. 1-13). https://doi.org/10.1145/3597503.3639191

[6] Gummadi, A. N., Napier, J. C., & Abdallah, M. (2024). XAI-IoT: an explainable AI framework for enhancing anomaly detection in IoT systems. IEEE Access, 12, 71024-71054. https://doi.org/10.1109/ACCESS.2024.3402446

[7] Jiao, T., Guo, C., Feng, X., Chen, Y., & Song, J. (2024). A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. Computers, Materials & Continua, 80(1). https://doi.org/10.32604/cmc.2024.053204

[8] Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., ... & Dou, D. (2022). Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. Knowledge and Information Systems, 64(12), 3197-3234. https://doi.org/10.1007/s10115-022-01756-8

[9] Au, J., Reid, D., & Bill, A. (2022, March). Challenges and opportunities of computer vision applications in aircraft landing gear. In 2022 IEEE aerospace conference (AERO) (pp. 1-10). IEEE. https://doi.org/10.1109/AERO53065.2022.9843684

[10] Kannan, V., Dao, D. V., & Li, H. (2023). An information fusion approach for increased reliability of condition monitoring with homogeneous and heterogeneous sensor systems. Structural Health Monitoring, 22(3), 1601-1612. https://doi.org/10.1177/1475921722111245

[11] Skinner, L. T., & Johnson, M. A. (2024, November). Bayesian networks for interpretable and extensible multisensor fusion. In Artificial Intelligence for Security and Defence Applications II (Vol. 13206, pp. 11-23). SPIE. https://doi.org/10.1117/12.3028532

[12] Dong, G., Tang, M., Wang, Z., Gao, J., Guo, S., Cai, L., ... & Boukhechba, M. (2023). Graph neural networks in IoT: A survey. ACM Transactions on Sensor Networks, 19(2), 1-50. https://doi.org/10.1145/3565973

[13] Wang, H., Li, J., McDonald, B. E., Farrell, T. R., Huang, X., & Clancy, E. A. (2023). Comparison between two time synchronization and data alignment methods for multi-channel wearable biosensor systems using BLE protocol. Sensors, 23(5), 2465. https://doi.org/10.3390/s23052465

[14] Tong, J., Liu, C., Zheng, J., & Pan, H. (2023). Multi-sensor information fusion and coordinate attention-based fault diagnosis method and its interpretability research. Engineering Applications of Artificial Intelligence, 124, 106614. https://doi.org/10.1016/j.engappai.2023.106614

[15] Li, L., Lv, M., Jia, Z., Jin, Q., Liu, M., Chen, L., & Ma, H. (2023). An effective infrared and visible image fusion approach via rolling guidance filtering and gradient saliency map. Remote Sensing, 15(10), 2486. https://doi.org/10.3390/rs15102486

[16] Chauhan, S., Vashishtha, G., & Zimroz, R. (2024). Analysing recent breakthroughs in fault diagnosis through sensor: a comprehensive overview. Computer Modeling in Engineering & Sciences, 141(3), 1983. https://doi.org/10.32604/cmes.2024.055633

[17] . Fang, Y., Min, H., Wu, X., Lei, X., Chen, S., Teixeira, R., & Zhao, X. (2023). Toward interpretability in fault diagnosis for autonomous vehicles: Interpretation of sensor data anomalies. IEEE Sensors Journal, 23(5), 5014-5027. https://doi.org/10.1109/JSEN.2023.3236838

[18] Wang, X., Wen, L., Cai, J., & Fang, K. (2024, July). Intelligent interference information fusion for security of uav forest remote sensing image detection. In Proceedings of the 15th Asia-Pacific Symposium on Internetware (pp. 313-316). https://doi.org/10.1145/3671016.3671384

[19] Anumula, S. R. (2022). Transparent and Auditable Decision-Making in Enterprise Platforms. International Journal of Research and Applied Innovations, 5(5), 7691-7702. https://doi.org/10.15662/IJRAI.2022.0505007

[20] Gawade, V., Zhang, B., & Guo, Y. (2023). Explainable AI for layer-wise emission prediction in laser fusion. CIRP Annals, 72(1), 437-440. https://doi.org/10.1016/j.cirp.2023.03.009

[21] Zhang, Y., Feng, K., Ma, H., Yu, K., Ren, Z., & Liu, Z. (2022). MMFNet: Multisensor data and multiscale feature fusion model for intelligent cross-domain machinery fault diagnosis. IEEE Transactions on Instrumentation and Measurement, 71, 1-11. https://doi.org/10.1109/TIM.2022.3213016

[22] Gahlan, N., & Sethia, D. (2024). Federated learning inspired privacy sensitive emotion recognition based on multi-modal physiological sensors. Cluster Computing, 27(3), 3179-3201. https://doi.org/10.1007/s10586-023-04133-4

[23] Wang, Y., Bai, X., Liu, C., & Tan, J. (2022). A multi-source data feature fusion and expert knowledge integration approach on lithium-ion battery anomaly detection. Journal of Electrochemical Energy Conversion and Storage, 19(2), 021003. https://doi.org/10.1115/1.4051716

[24] Brödermann, T., Bruggemann, D., Sakaridis, C., Ta, K., Liagouris, O., Corkill, J., & Van Gool, L. (2024, September). Muses: The multi-sensor semantic perception dataset for driving under uncertainty. In European Conference on Computer Vision (pp. 21-38). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-73202-7_2

[25] Lin, X., Chao, S., Yan, D., Guo, L., Liu, Y., & Li, L. (2023). Multi-sensor data fusion method based on self-attention mechanism. Applied Sciences, 13(21), 11992. https://doi.org/10.3390/app132111992

[26] Cao, Y., Wang, N., Xiao, C., Yang, D., Fang, J., Yang, R., ... & Li, B. (2021, May). Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In 2021 IEEE symposium on security and privacy (SP) (pp. 176-194). IEEE. https://doi.org/10.1109/SP40001.2021.00076

[27] Ignatious, H. A., El-Sayed, H., Khan, M. A., & Kulkarni, P. (2023). A generic framework for enhancing autonomous driving accuracy through multimodal data fusion. Applied Sciences, 13(19), 10749. https://doi.org/10.3390/app131910749

[28] Sinha, A., & Das, D. (2023). XAI-LCS: Explainable AI-based fault diagnosis of low-cost sensors. IEEE Sensors Letters, 7(12), 1-4. https://doi.org/10.1109/LSENS.2023.3330046

[29] Chen, S., & Wang, J. (2024). MultiCogniGraph: A multimodal data fusion and graph convolutional network-based multi-hop reasoning method for large equipment fault diagnosis. Computational Intelligence, 40(3), e12646. https://doi.org/10.1111/coin.12646

[30] Geng, H., Liu, H., Ma, L., & Yi, X. (2021). Multi-sensor filtering fusion meets censored measurements under a constrained network environment: advances, challenges and prospects. International Journal of Systems Science, 52(16), 3410-3436. https://doi.org/10.1080/00207721.2021.2005178

[31] Liu, G., Jiang, Y., Zhong, K., Yang, Y., & Wang, Y. (2023). A time series model adapted to multiple environments for recirculating aquaculture systems. Aquaculture, 567, 739284. https://doi.org/10.1016/j.aquaculture.2023.739284

[32] Perez-Cerrolaza, J., Abella, J., Borg, M., Donzella, C., Cerquides, J., Cazorla, F. J., ... & Flores, J. L. (2024). Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. ACM Computing Surveys, 56(7), 1-40.https://doi.org/10.1145/3626314

[33] Yuan, L., Andrews, J., Mu, H., Vakil, A., Ewing, R., Blasch, E., & Li, J. (2022). Interpretable passive multi-modal sensor fusion for human identification and activity recognition. Sensors, 22(15), 5787. https://doi.org/10.3390/s22155787

[34] Terziyan, V., & Vitko, O. (2022). Explainable AI for industry 4.0: semantic representation of deep learning models. Procedia Computer Science, 200, 216-226. https://doi.org/10.1016/j.procs.2022.01.220

[35] Wang, X., Wang, B., Wu, Y., Ning, Z., Guo, S., & Yu, F. R. (2024). A survey on trustworthy edge intelligence: From security and reliability to transparency and sustainability. IEEE Communications Surveys & Tutorials, 27(3), 1729-1757. https://doi.org/10.1109/COMST.2024.3446585