JURNAL AR RO'IS MANDALIKA (ARMADA)
Journal website: https://ojs.cahayamandalika.com/index.php/armada

ISSN: 2774-8499 Vol. 6 No. 2 (2026)

Research Article

Exploring Multimodal AI Frameworks for Real-Time Decision Making in Edge Devices

Darmin[1], Imam Taufik[2], Miswadi[3], Kustiyono[4], Sahlan M. Saleh[5]

Institut Sains dan Teknologi Alkamal (ISTA), Indonesia [1]
Universitas Kahuripan Kediri, Indonesia [2]
Politeknik Meta Industri Cikarang, Indonesia [3]
Universitas Ngudi Waluyo, Indonesia [4]
Universitas Yapis Papua, Indonesia [5]
Corresponding Author, Email: darmin@ista.ac.id

Abstract

The rapid advancement of Artificial Intelligence (AI) and edge computing has driven the demand for intelligent systems capable of real-time decision making under limited computational resources. In particular, multimodal AI, which integrates heterogeneous data sources such as visual, audio, and sensor signals, plays a crucial role in enhancing contextual awareness and decision accuracy at the edge. This study aims to explore and conceptualize a multimodal AI framework that supports real-time decision making on edge devices while addressing challenges related to resource constraints, data privacy, and decision transparency. The research adopts a qualitative literature review approach, employing a Systematic Literature Review (SLR) method to analyze relevant studies published between 2018 and 2025. Data were collected from reputable academic databases and analyzed using thematic content analysis to identify key architectural components, fusion strategies, optimization techniques, and privacy-preserving mechanisms. The findings indicate that hybrid multimodal fusion, combined with model compression, dynamic inference, and federated learning, significantly improves efficiency, privacy protection, and explainability in edge-based AI systems. This study contributes a comprehensive conceptual framework that can guide future development and deployment of adaptive, efficient, and trustworthy multimodal AI solutions for real-time edge intelligence applications.

This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0).

Keywords: Multimodal AI, Edge Computing, Real-Time Decision Making.

INTRODUCTION

The rapid development of Artificial Intelligence (AI) technologies has significantly transformed the way computational systems process, analyze, and make real-time decisions across various application domains. The integration of AI with edge computing enables data processing directly on edge devices, thereby reducing latency, improving energy efficiency, and strengthening user privacy (Prabaharan et al., 2025). In this context, multimodal AI—which combines diverse data sources such as visual, textual, and sensory inputs—forms a crucial foundation for adaptive and intelligent decision-making (Hussain et al., 2024). The demand for systems capable of efficiently processing multimodal data on edge devices continues to grow in parallel with the expansion of IoT applications and modern autonomous systems (Yuan, 2024).

The paradigm shift from cloud computing to edge computing is driven by the need for rapid response times and bandwidth efficiency in environments with limited connectivity (Ficili et al., 2025). AI integration at the edge allows for data-driven decision-making directly at the data source, reducing dependency on centralized data centers (Ficili et al., 2025). This is critical in applications such as autonomous vehicles, intelligent healthcare, and Industry 4.0 infrastructures that demand high reliability and minimal response times (Ramesh & Praveen, 2021). The combination of multimodal AI and edge computing promises to enhance systems' adaptability to dynamic real-world contexts (Pascher, 2024).

However, implementing multimodal AI on edge devices still faces major challenges such as limited computational capacity, energy consumption, and the complexity of fusing heterogeneous data modalities (Huang et al., 2025). Traditional cloud-based systems often fail to meet real-time and resource efficiency requirements (Uddin, 2025). Therefore, it is necessary to develop frameworks that optimize workload distribution between the edge and the cloud, ensure data privacy, and maintain model accuracy in diverse environments (Rjoub et al., 2025). Innovations in federated learning and lightweight model optimization present potential solutions to these challenges (Rajesh, 2025).

The urgency of this research lies in the growing need for AI systems capable of adaptive, accurate, and transparent decision-making directly on edge devices without compromising efficiency or data security. Such systems are vital in time-critical domains such as autonomous driving, healthcare, and smart industrial networks (AlNusif, 2025). By developing a distributed multimodal AI framework, this research aims to contribute to the realization of adaptive, sustainable, and trustworthy Edge Intelligence.

Previous studies have demonstrated significant progress in AI implementation on edge systems and multimodal processing. For instance, (Jiang et al., 2025) developed *Farm-LightSeek*, an edge-centric agricultural IoT data analytics framework that integrates lightweight LLMs for real-time decision-making, while (Wang, 2025) highlighted the potential of multimodal AI systems in supporting medical decision-making using big data. Moreover, (Zhou et al., 2019) emphasized the synergy between AI and edge computing for fast decision support in electrical systems. However, these studies remain domain-specific and do not yet propose a generalized framework that can be applied across multiple sectors for real-time multimodal AI-based decision-making.

Based on this background, the present study aims to explore and design a multimodal AI framework capable of integrating diverse data sources to support real-time decision-making on edge devices. It also seeks to identify algorithmic and architectural optimization approaches that enhance resource efficiency, data privacy, and decision transparency. Therefore, the findings of this study are expected to make both theoretical and practical contributions to the advancement of multimodal AI-based Edge Intelligence across various industrial and societal applications.

METHOD

Type of Research

This research is a qualitative literature study, aimed at exploring, analyzing, and synthesizing theories, frameworks, and previous studies related to multimodal AI frameworks for real-time decision-making on edge devices. A literature study is appropriate because it allows for an in-depth understanding of a research topic through critical interpretation of existing scholarly works without direct empirical data collection (Creswell & Clark, 2017). The focus of this study is to identify conceptual relationships, technological trends, and research gaps regarding the integration of multimodal AI and edge computing in real-time decision systems.

Data Sources

The data sources in this study consist of secondary data derived from reputable academic publications such as international journal articles, conference proceedings, scholarly books, and technical research reports in the fields of Artificial Intelligence (AI), Edge Computing, and Multimodal Learning. The data selection process follows purposive sampling, prioritizing works based on relevance, credibility, and recency. Only articles published between 2018 and 2025 were included to ensure the representation of current advancements (Snyder, 2019). Databases such as Scopus, IEEE Xplore, ScienceDirect, and SpringerLink were used for data retrieval. Inclusion criteria comprised studies addressing multimodal AI integration, real-time decision-making, and edge computing architectures, while exclusion criteria eliminated publications lacking empirical or conceptual depth in these areas.

Data Collection Techniques

The data collection process employed a systematic documentation method using the Systematic Literature Review (SLR) framework as described by (Kitchenham & Charters, 2007). This method was implemented in four key stages:

1. Formulating research questions, focusing on how multimodal AI frameworks are designed and optimized for real-time decision-making at the edge.
2. Conducting a systematic search using keywords such as “multimodal AI framework,” “real-time decision-making,” “edge computing,” and “edge intelligence.”
3. Selecting studies based on inclusion and exclusion criteria to ensure relevance and quality.
4. Extracting key data and information from selected studies, including details about architecture, algorithms, challenges, and performance outcomes.

During the process, reference management tools such as Mendeley and Zotero were used to organize documents, manage citations, and avoid duplication. The systematic approach enhances transparency and replicability, ensuring that the review is both comprehensive and methodologically sound.

Data Analysis Methods

The collected data were analyzed using Thematic Content Analysis (TCA) as outlined by (Braun & Clarke, 2021). This method involves identifying, analyzing, and reporting patterns (themes) within the literature. The analysis process followed several stages:

1. Familiarization – reading and understanding all selected documents to grasp their contexts and main arguments.
2. Coding – labeling key concepts related to multimodal data fusion, model architectures, and edge decision mechanisms.
3. Theme development – grouping similar codes into broader analytical themes (e.g., resource efficiency, data privacy, decision transparency).
4. Interpretation and synthesis – integrating the identified themes into a conceptual framework that explains the interrelationship between multimodal AI, real-time decision-making, and edge computing.

This method enables the researcher to synthesize diverse findings into a coherent narrative and identify research gaps for future exploration (Nowell et al., 2017).

RESULT AND DISCUSSION

Multimodal AI Framework Design for Edge-Based Real-Time Decision Making

A comprehensive multimodal AI framework for real-time decision-making in edge devices can be conceptualized as a layered architecture composed of:

1. Data Acquisition Layer, integrating multimodal sensors such as cameras, microphones, and accelerometers;
2. Feature Encoding Layer, applying modality-specific neural encoders;
3. Fusion Layer, combining the extracted features;
4. Decision Layer, performing inference and generating context-aware actions; and
5. Feedback Loop, enabling adaptive learning and policy refinement.

At the data acquisition level, multimodal sources capture complementary information. For instance, visual sensors provide spatial context, while audio and tactile sensors capture temporal or environmental cues. Research shows that multimodal data improves reliability by compensating for noise or failure in individual modalities (Ngiam et al., 2011).

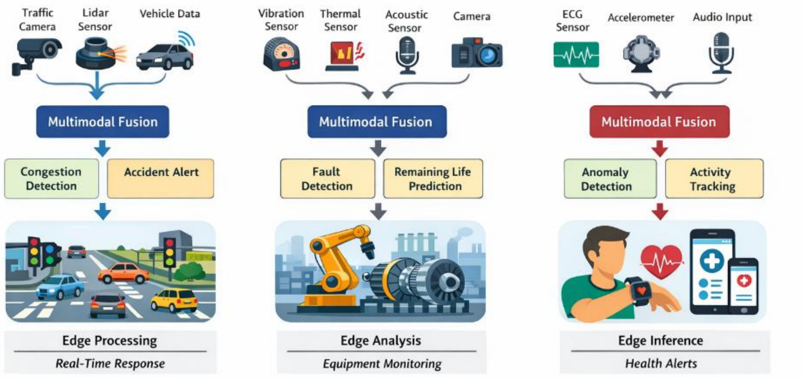

Figure 1. Multimodal AI for Real-Time Edge Decision Making

At the feature encoding layer, different neural architectures are used depending on modality characteristics:

1. CNNs are effective for extracting spatial features from images and video.
2. RNNs or Transformers capture time-varying dependencies in audio and language signals.
3. Temporal convolution or probabilistic models are used for motion and sensor streams.

Studies have validated that modality-specific encoders significantly enhance overall accuracy and robustness in multimodal processing compared to unified feature extraction methods (Baltrušaitis et al., 2018).

1. Fusion Strategies for Multimodal Decision Making

One of the central design decisions involves selecting the appropriate fusion strategy. There are three main categories:

a. Early Fusion

Early fusion integrates raw or low-level features before processing. It offers strong interaction modeling but suffers from sensitivity to misalignment and data heterogeneity. For example, combining raw EEG and video signals yields improved emotion recognition accuracy but requires strict synchronization (Zheng et al., 2014).

b. Late Fusion

Late fusion aggregates decisions from modality-specific predictors. It provides modularity and robustness when modalities are missing or degraded. However, it may not optimally leverage cross-modal correlations (Atrey et al., 2010).

c. Hybrid Fusion

Hybrid fusion is the most effective approach in edge environments, balancing performance and flexibility. Intermediate feature interactions are combined with decision-level aggregation, enabling devices to dynamically adjust computational cost, latency, and accuracy. Empirical studies show hybrid fusion improves performance in dynamic settings such as wearable monitoring and autonomous driving (Akhtar & Mian, 2018).

This capability is essential for real-time decision-making in edge devices, where network availability and power consumption vary unpredictably.

A key real-world example is edge-based smart traffic management. Traffic cameras, roadside sensors, and vehicle telematics provide multimodal data used to detect congestion, violations, and accidents. On-edge fusion enables real-time response, such as adaptive traffic signals. Edge processing reduces latency and network dependency, providing millisecond-level reaction time, which is critical for dynamic traffic flow optimization (Shi et al., 2016). Studies demonstrate that combining vision data, radar signals, and trajectory data enables more accurate hazard detection than using single modalities (Zhao et al., 2024). Edge processing further reduces data transmission costs and enhances privacy by avoiding raw video uploads.

Another example is predictive maintenance in Industry 4.0. Multimodal signals—such as vibration, temperature, acoustic emissions, and image data—are fused on edge devices to predict equipment failures. Multimodal fusion improves fault detection accuracy compared to single-sensor approaches, especially when monitoring complex industrial processes (Ma et al., 2022). Edge computing enables faster anomaly detection, minimizing downtime. Industrial edge inference systems have demonstrated fault detection accuracy improvements of up to 20–30% when using multimodal rather than single-modal approaches (Tao et al., 2018).

In healthcare, multimodal fusion enables monitoring systems to combine biosignals (ECG, accelerometer) and contextual data (speech emotion, activity recognition). Edge-based inference allows real-time abnormality detection without uploading sensitive patient data to cloud servers. Studies in remote patient monitoring confirm multimodal models outperform single-biometric systems for early disease detection (Nweke et al., 2018). Edge devices support continuous monitoring with minimal latency, which is essential for emergency response (Lane et al., 2015).

Algorithmic Optimization for Resource Efficiency

Edge devices such as smartphones, IoT sensors, wearables, and embedded controllers inherently possess limited compute power, memory, and energy budget compared to cloud servers. Running multimodal AI models directly on these devices challenges traditional deep learning architectures, which often assume abundant computational resources. Without specialized optimization, multimodal models can suffer from high latency, excessive power consumption, and poor real-time responsiveness—making them impractical for real-time decision-making at the edge.

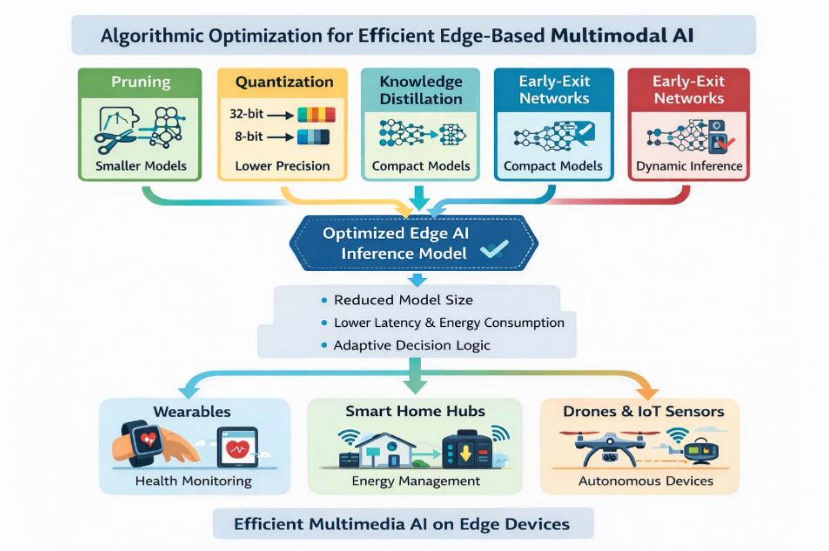

Figure 2. Algorithmic Optimization for Edge-Based Multimodal AI

To overcome these constraints, the framework must implement resource-efficient algorithmic strategies that retain model accuracy while reducing computational overhead.

1. Model Compression Techniques

   Model compression refers to a set of methods that reduce the size and complexity of neural networks without significantly degrading performance. The three most impactful strategies in edge optimization are:

   a. Pruning

      Pruning removes redundant or less important weights from neural networks, resulting in sparse models. Research shows that pruning can reduce model size by up to 90% in some cases while maintaining accuracy (Han et al., 2015). In multimodal AI, pruning must be modality-aware—safely removing weights in both visual and sensor encoders without degrading feature extraction quality. For example, researchers successfully pruned multimodal speech and facial processing models for smart assistants on embedded platforms.

b. Quantization

Quantization reduces the precision of weights and activations from 32-bit floating point to lower bit-width formats (e.g., 8-bit integer). Quantized models significantly reduce memory footprint and demand less energy during inference (Jacob et al., 2018). Studies show 8-bit quantization can reduce inference latency by 30–60% with minimal accuracy loss for vision and sensor models. For multimodal systems, quantization must consider cross-modal representation stability to avoid disproportionate performance drops.

c. Knowledge Distillation

In knowledge distillation, a large “teacher” model trains a smaller “student” model to mimic its outputs. Distilled models achieve similar performance to large models while being more compact (Hinton et al., 2015). This strategy is critical in multimodal fusion where models must balance complex feature interactions while remaining lightweight.

2. Early-Exit and Dynamic Inference Architectures

Beyond compression, dynamic inference techniques adapt model complexity at run-time based on input difficulty:

a. Early-Exit Architectures

Early-exit models allow inference to stop before full model propagation once a confident decision is achieved at an intermediate layer. For edge devices, early-exit significantly reduces average latency and energy usage because “easy” inputs require less computation (Teerapittayanon et al., 2016).

b. Dynamic Modality Selection

In multimodal frameworks, not all modalities are equally informative at all times. Dynamic selection chooses only the most relevant modalities in real time, reducing processing cost.

For example:

1) A smart surveillance system may skip audio processing when visual cues already indicate a critical event.

2) A wearable device may bypass language processing when only biosignals are relevant.

Dynamic architectures are essential for context-aware resource efficiency.

3. Real-World Case 1: Wearable Healthcare Monitoring

Wearable health devices (e.g., smartwatches with ECG, accelerometer, and SpO2 sensors) require continuous monitoring while preserving battery life. Multimodal AI fuses sensor streams to detect abnormal conditions such as arrhythmia or falls.

Optimization Strategies Applied:

a. Quantized CNN and LSTM models process combined biosignal and movement data.
b. Early-exit layers allow immediate classification when patterns are highly characteristic (e.g., clear arrhythmia waveform).
c. Knowledge distillation produces compact models deployable on microcontrollers.

These optimizations enable 24/7 monitoring with minimal energy draw, ensuring real-time alerts without cloud dependency.

Architectural Strategies for Data Privacy Preservation

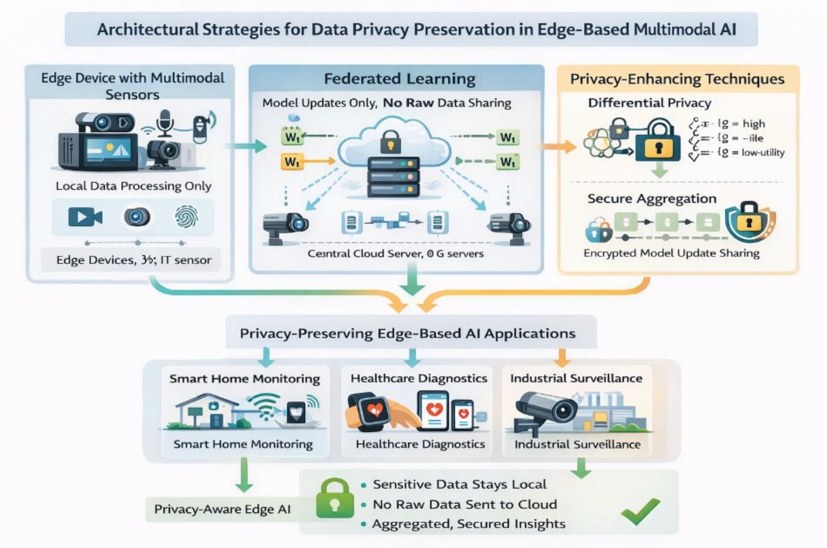
Architectural Strategies for Data Privacy Preservation in Edge-Based Multimodal AI
Edge Device with Multimodal Sensors
Local Data Processing Only
Edge Devices, 3½; IT sensor
Federated Learning
Model Updates Only, No Raw Data Sharing
W₁
W₁
W₁
W₁
Central Cloud Server, 0 G servers
Privacy-Enhancing Techniques
Differential Privacy
{g = high
{g = –iile
{g = low-utility
Secure Aggregation
Encrypted Model Update Sharing
Privacy-Preserving Edge-Based AI Applications
Smart Home Monitoring
Healthcare Diagnostics
Industrial Surveillance
Smart Home Monitoring
Healthcare Diagnostics
Industrial Surveillance
Privacy-Aware Edge AI
• Sensitive Data Stays Local
• No Raw Data Sent to Cloud
• Aggregated, Secured Insights

Figure 3. Privacy-Preserving Multimodal AI Architecture at the Edge

1. Edge-Centric Processing for Minimizing Privacy Risks

A core architectural strategy for privacy preservation in multimodal AI on edge devices is to shift computation physically closer to the data source. Edge AI processes sensitive data such as images, voice, or biometric signals locally on the device or nearby edge server, rather than transmitting raw data to centralized cloud servers. This approach intrinsically reduces privacy risks associated with network interception, unauthorized access, or regulatory violations (e.g., GDPR) because raw personal data never leaves the edge node. Local processing also limits the exposure of identifiable information, keeping only high-level feature representations or model updates for further aggregation or analysis.

In this context, methods such as differential privacy and secure aggregation can be combined with edge processing to further strengthen privacy. Differential privacy adds controlled noise to model updates, making it statistically improbable to reconstruct individual data points from aggregate information. Secure aggregation protocols cryptographically ensure that model updates from multiple devices remain confidential during federated learning rounds.

2. Federated Learning: Collaborative Model Training Without Raw Data Sharing

Federated Learning (FL) has emerged as a foundational privacy-preserving technique in edge architectures. In FL, each edge device independently trains a local model on its own data and subsequently sends only model updates (e.g., gradients or weights) to a centralized coordinator or aggregator. This means sensitive user data—such as multimodal sensor readings or video feeds—never leave the device, significantly mitigating privacy and confidentiality concerns in distributed AI systems.

In multimodal scenarios, FL can enable collaborative learning across devices equipped with diverse sensor types (e.g., cameras, microphones, wearables), allowing generalized models to emerge without centralizing the underlying raw data. A systematic survey indicates that FL not only preserves privacy but also supports collaborative learning on heterogeneous edge datasets, although challenges such as model heterogeneity and non-IID (non-identically and independently distributed) data must be addressed.

3. Real-World Case: Privacy-Preserving Medical Image Classification

One compelling real-world example is the deployment of FL in medical image classification across multi-institutional edge servers. In this application, hospitals or medical centers perform local training on sensitive patient images (e.g., skin lesion images) while participating in a global model training process. Through federated edge collaborations, each institution retains control of its data, ensuring compliance with strict privacy regulations such as HIPAA or GDPR. The global model benefits from the varied dataset without exposing individual patient records to external entities.

This architecture demonstrated that edge-based collaborative training can achieve accuracy and reliability comparable to centralized training while preserving privacy and reducing the need for large centralized data repositories. Such configurations are especially relevant for healthcare scenarios requiring high levels of confidentiality and legal compliance.

4. Real-World Case: Privacy-Preserving Multimodal Health Diagnostics

Another case is the application of federated learning for multimodal health prediction, such as adaptive menstrual health tracking using multimodal biosensing (e.g., radar sensing, physiological sensors). In this framework, multimodal physiological protocols run locally on user devices. Only model parameters or encrypted representations are shared for collaborative learning, ensuring sensitive health data remains private and localized.

This approach avoids centralizing private health records, enabling AI features such as personalized inference, continuous model improvement, and cross-individual pattern learning without exposing raw health signals.

5. Complementary Privacy Techniques: Differential Privacy & Secure Aggregation

While FL limits raw data sharing, advanced privacy techniques further enhance protection:

a. Differential Privacy (DP): Introduces carefully calibrated noise to model updates to protect against inference attacks (e.g., membership inference) while balancing accuracy. Applying DP during FL can preserve individual privacy more robustly.
b. Secure Aggregation: Cryptographic protocols aggregate model updates in a way that individual contributions remain encrypted, ensuring the coordinator cannot access individual updates.

Together, these methods build a secure pipeline where on-device pre-processing, model training, and parameter sharing preserve privacy without forfeiting collaborative learning benefits.

Enhancing Decision Transparency and Explainability

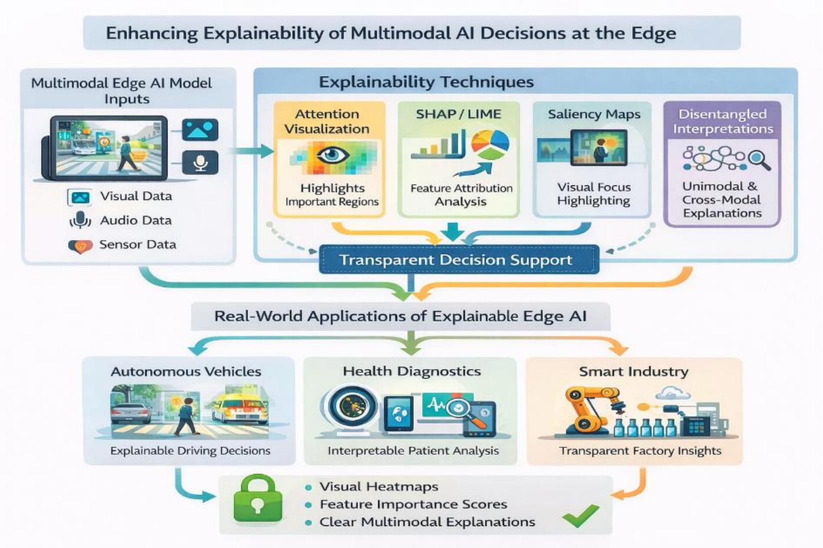
Enhancing Explainability of Multimodal AI Decisions at the Edge
Multimodal Edge AI Model Inputs
Visual Data
Audio Data
Sensor Data
Explainability Techniques
Attention Visualization
Highlights Important Regions
SHAP / LIME
Feature Attribution Analysis
Saliency Maps
Visual Focus Highlighting
Disentangled Interpretations
Unimodal & Cross-Modal Explanations
Transparent Decision Support
Real-World Applications of Explainable Edge AI
Autonomous Vehicles
Explainable Driving Decisions
Health Diagnostics
Interpretable Patient Analysis
Smart Industry
Transparent Factory Insights
Visual Heatmaps
Feature Importance Scores
Clear Multimodal Explanations

Figure 4. Explainable Multimodal AI at the Edge

1. The Importance of Explainability in Edge-Deployed Multimodal AI

   As multimodal AI systems become integral to real-time decision-making—especially on resource-limited edge devices—explainability and transparency are no longer optional features but essential requirements. Edge AI often supports decisions in safety-critical domains such as autonomous driving, healthcare diagnostics, and industrial automation. In these contexts, stakeholders need to understand not only what the model predicts but why it arrives at a particular decision. Without transparency, complex models act as "black boxes," which can undermine trust, complicate debugging, hamper compliance with regulatory standards (e.g., GDPR or medical safety regulations), and reduce accountability. Explainable AI (XAI) addresses these concerns by providing insights into a model's internal reasoning pathways and decision logic, thereby improving interpretability for developers and end users alike.

2. Key Explainability Techniques for Multimodal Edge AI

   Explainable multimodal models use a variety of techniques to make decisions understandable:

   a. Attention Mechanisms and Attention Visualization

      Modern multimodal networks often employ attention mechanisms (derived from transformer architectures) to weight the importance of data features across modalities. Visualizing attention scores highlights which parts of an input (e.g., image regions, text tokens, sensor segments) the model focused on during inference. Such attention maps serve as intuitive explanation tools that show what contributes most to a prediction, thereby fostering transparency. Examples include cross-modal attention where, for instance, text description may highlight specific image regions that inform a classification outcome.

b. Feature Attribution and SHAP/LIME

Model-agnostic post-hoc explainability methods such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) assign importance scores to features by simulating their absence and measuring changes in output. These methods can be applied not only to tabular or visual inputs but also extended to multimodal settings to quantify how each modality contributed to the final decision, enabling users to understand contributions from different data types.

c. Saliency Maps and Activation Visualization

For vision-based modalities, saliency maps and class activation mappings (e.g., Grad-CAM) highlight image regions that heavily influence the model's prediction. Saliency maps can be visualized even on edge devices and provide a spatial explanation of decision focus, which is critical in applications like medical image analysis or object detection in autonomous systems.

d. Multimodal Disentangled Interpretations

Recent research such as DIME (Disentangled Local Explanations) emphasizes fine-grained interpretability by separating unimodal contributions from multimodal interactions, allowing stakeholders to see not just which features mattered, but how modalities interacted to influence decisions.

3. Real-World Case: Transparent Decision Support in Autonomous Driving

Autonomous vehicles (AVs) must operate safely in dynamic environments where decisions impact human lives. Explainability is crucial in AV systems to justify actions like braking, lane changes, or pedestrian avoidance. In explainable autonomous driving models, attention visualization combined with semantic explanations help engineers and regulators interpret model decisions. For instance, attention maps can show that a vehicle's braking decision was influenced by a pedestrian's movement pattern and road signage context—rather than irrelevant background features. Researchers have developed explainable frameworks that jointly generate both visual attention maps and textual rationales for AV decisions, enhancing interpretability and debugging capability for safety validation.

4. Real-World Case: Interpretable Multimodal Health Diagnostics

   In healthcare diagnostics, explainability is indispensable. Consider a multimodal AI model used to diagnose neurological disorders by integrating imaging data (e.g., MRI) with clinical text reports and physiological sensor signals. Without XAI methods like attention visualization or feature attribution, clinicians cannot trust or validate AI-assisted recommendations. For example, explainable multimodal diagnostics systems can visually highlight anomalous brain regions on MRI (via saliency or activation maps) while simultaneously indicating which aspects of clinical history contributed most to a diagnosis. This combination of visual and textual explanation supports clinicians in making informed decisions, thereby improving diagnostic confidence and patient outcomes.

5. Integrating Explainability with Edge Constraints

   While XAI methods traditionally introduce computational overhead, recent research and frameworks demonstrate that lightweight explainability techniques can be adapted to edge environments. For instance, LIME and saliency maps have been benchmarked on constrained devices and shown to offer acceptable latency and interpretability trade-offs. LIME tends to provide a balanced explanation quality with manageable resource use, while saliency maps offer fast visual insights with minimal overhead, making them suitable for real-time edge AI inference in latency-critical applications.

CONCLUSION

This study demonstrates that multimodal AI frameworks integrated with edge computing offer a promising solution for real-time decision making in resource-constrained environments. By combining multiple data modalities, edge-based systems achieve higher robustness, contextual understanding, and decision accuracy compared to single-modal approaches. The analysis highlights that hybrid fusion strategies, lightweight model optimization, federated learning, and explainable AI techniques are critical components for achieving efficiency, privacy preservation, and transparency.

Overall, multimodal edge intelligence represents a strategic advancement for time-sensitive and privacy-critical applications such as healthcare monitoring, intelligent transportation, and industrial automation.

Practical Implications

From a practical perspective, this framework can guide system designers and practitioners in deploying AI models directly on edge devices with minimal latency and reduced reliance on cloud infrastructure. The integration of optimization and privacy-preserving mechanisms enables organizations to improve operational efficiency while maintaining compliance with data protection regulations.

Future Research Directions

Future studies should focus on empirical validation of the proposed framework through prototyping and benchmarking on real edge platforms. Further research is also recommended to explore adaptive multimodal learning under dynamic environments and to enhance explainability techniques suitable for ultra-low-power edge devices.

Bibliography

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, *6*, 14410–14430.

AlNusif, M. (2025). Explainable AI in edge devices: A lightweight framework for real-time decision transparency. *Int. J. Eng. Comput. Sci*, *14*(07), 27447–27472.

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, *16*(6), 345–379.

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443.

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Sage publications.

Ficili, I., Giacobbe, M., Tricomi, G., & Puliafito, A. (2025). From sensors to data intelligence: Leveraging IoT, cloud, and edge computing with AI. *Sensors*, *25*(6), 1763.

Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural network. *Advances in Neural Information Processing Systems*, *28*.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *ArXiv Preprint ArXiv:1503.02531*.

Huang, X., Wang, H., Shiyin, Q., & Su-Kit, T. (2025). Embedded artificial intelligence: A comprehensive literature review. *Electronics*, *14*(17), 3468.

Hussain, M., O'Nils, M., Lundgren, J., & Mousavirad, S. J. (2024). A comprehensive review on deep learning-based data fusion. *IEEE Access*.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2704–2713.

Jiang, D., Shen, Z., Zheng, Q., Zhang, T., Xiang, W., & Jin, J. (2025). Farm-LightSeek: An Edge-centric Multimodal Agricultural IoT Data Analytics Framework with Lightweight LLMs. *IEEE Internet of Things Magazine*.

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*.

Lane, N. D., Bhattacharya, S., Georgiev, P., Forlivesi, C., & Kawsar, F. (2015). An early resource characterization of deep learning on wearables, smartphones and internet-of-things devices. *Proceedings of the 2015 International Workshop on Internet of Things towards Applications*, 7–12.

Ma, Y., Wen, G., Cheng, S., He, X., & Mei, S. (2022). Multimodal convolutional neural network model with information fusion for intelligent fault diagnosis in rotating machinery. *Measurement Science and Technology*, *33*(12), 125109.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. *ICML*, *11*, 689–696.

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, *16*(1), 1609406917733847.

Nweke, H. F., Teh, Y. W., Al-Garadi, M. A., & Alo, U. R. (2018). Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges. *Expert Systems with Applications*, *105*, 233–261.

Pascher, M. (2024). *An Interaction Design for AI-enhanced Assistive Human-Robot Collaboration*. Dissertation, Duisburg, Essen, Universität Duisburg-Essen, 2024.

Prabaharan, G., Vidhya, S., Chithrakumar, T., Sika, K., & Balakrishnan, M. (2025). AI-driven computational frameworks: Advancing edge intelligence and smart systems. *International Journal of Computational and Experimental Science and Engineering*, *11*(1), 1363–1369.

Rajesh, M. (2025). Adaptive Edge-Federated AI Framework for Contactless Menstrual Health Prediction Using Multimodal Physiological Intelligence. *MethodsX*, *15*, 103665.

Ramesh, G., & Praveen, J. (2021). Artificial intelligence (ai) framework for multi-modal learning and decision making towards autonomous and electric vehicles. *E3S Web of Conferences*, *309*, 1167.

Rjoub, G., Elmekki, H., Islam, S., Bentahar, J., & Dssouli, R. (2025). A hybrid swarm intelligence approach for optimizing Multimodal Large Language Models deployment in edge-cloud-based Federated Learning environments. *Computer Communications*, *237*, 108152.

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, *3*(5), 637–646.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339.

Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., & Sui, F. (2018). Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology*, *94*(9), 3563–3576.

Teerapittayanon, S., McDanel, B., & Kung, H.-T. (2016). Branchynet: Fast inference via early exiting from deep neural networks. *2016 23rd International Conference on Pattern Recognition (ICPR)*, 2464–2469.

Uddin, B. (2025). Development of Intelligent Internet of Things Systems for Real-Time Data Processing and Decision Making. *International Journal for Science Review*, *2*(6).

Wang, H. (2025). Research on Medical Intelligent Decision Support System Integrating Multimodal AI Big Data. *SHS Web of Conferences*, *213*, 2033.

Yuan, P. (2024). Artificial intelligence in the Internet of Things: Integrating and optimizing AI algorithms for real-time data processing and decision-making. *Applied and Computational Engineering*, *102*(1), 84–89.

Zhao, F., Zhang, C., & Geng, B. (2024). Deep multimodal data fusion. *ACM Computing Surveys*, *56*(9), 1–36.

Zheng, W.-L., Dong, B.-N., & Lu, B.-L. (2014). Multimodal emotion recognition using EEG and eye tracking data. *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5040–5043.

Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, *107*(8), 1738–1762.