Multimodal Cognitive Fusion for Low-Resource Environments

Khushi Rathore[/], Harsha Narayan[/], Shashwat Singh[/],

Sowmya Natarajan[0000-0002-9888-6078]

Dept of Electronics & Communication Engineering S.R.M. Institute of Science & Technology Chennai, India.

sowmyan1@srmist.edu.in

Abstract. The growing use of Artificial Intelligence in real-world systems has created a strong need for models that can understand and process different types of data at the same time. However, current multimodal fusion models rely heavily on large datasets and computing power, which makes them unsuitable for low-resource environments. This paper presents CAF-Net (Cognitive Adaptive Fusion Network), a lightweight and resource-efficient framework inspired by human thinking. CAF-Net uses specific encoders for each modality, a Cognitive Attention Fusion Layer (CAFL), and an Adaptive Decision Layer (ADL) to balance accuracy and efficiency based on system limits. The CAFL dynamically assigns attention weights to the most useful modalities, while the ADL adjusts inference by turning modalities on or off depending on real-time resource availability. Experimental results show that CAF-Net achieves accuracy similar to transformer-based models while significantly lowering computing demands, energy use, and latency. Its adaptability makes it ideal for edge AI, IoT healthcare, and mobile computing applications in settings with limited resources.

Keywords: Multimodal Learning, Cognitive Fusion, Low-Resource AI, Edge Computing, TinyML

1 Introduction

In today's world, artificial intelligence (AI) has become a crucial part of many areas of human life, including healthcare, transportation, agriculture, education, and defence. The growing use of AI in real systems has led to smarter automation, better decision-making, and human-like perception. This makes it one of the most transformative technologies of the 21st century. A significant development driving this change is multimodal learning. This area allows AI models to process and combine information from various sources like text, images, audio, and sensor data. This capability helps machines understand complex environments more fully, similar to how the human brain mixes visual, auditory, and contextual information to make sense of the world [1, 2].

For example, an AI healthcare diagnostic system may analyse medical images, doctors' notes, and physiological sensor data at the same time to produce more accurate predictions. In another case, a surveillance system can combine video, sound, and

Research Paper
DOI: https://doi.org/10.46793/BISEC25.405R
Part of ISBN: 978-86-89755-40-4

sensor inputs to detect threats in changing environments. This multimodal intelligence improves understanding and decision-making, allowing AI systems to work in various real-world situations. However, traditional multimodal systems, like vision-language transformers and large fusion models, often require a lot of computing power and memory. They depend on large datasets and powerful hardware like GPUs and TPUs. This need for high computational power and a stable connection makes them unsuitable for low-resource environments (LREs), where energy, processing capacity, memory, and bandwidth are limited. In these situations, using large multimodal models can lead to increased delays, high energy use, and poor scalability, making them impractical for real-time or mobile applications. The gap between AI capabilities and available resources highlights the urgent need for adaptable AI frameworks that can work smartly even when resources are scarce.

Low-resource environments are defined by limited computing power, restricted memory, unstable network connections, and a lack of energy. These conditions are common in developing areas, rural settings, mobile devices, and edge computing environments where access to cloud computing and large data centres is restricted. Examples include remote healthcare systems that use IoT-based wearable devices with small batteries, autonomous drones that monitor crops or conduct surveillance in disaster zones, smartphones that offer AI services in rural areas with poor internet, and embedded edge devices used in agriculture, defence, or environmental monitoring, where local processing is necessary without cloud support. In these applications, AI systems need to deliver accurate results while also being efficient, ensuring low delays and power consumption. Finding the right balance between accuracy, energy efficiency, and computing costs is one of the biggest challenges in using AI in resource-limited operations [3].

Traditional multimodal models typically rely on fast cloud infrastructure for both training and inference. However, this cloud dependence is not always practical or sustainable, especially when handling sensitive data or unstable connections. Continuous transferring of multimodal data, like high-resolution images and sensor streams, to cloud servers adds delays and increases energy use, which defeats the purpose of needing real-time responses. This has led to a demand for AI models that can work at the edge, processing and analysing data locally. These systems should be small enough to fit into embedded hardware while also being smart enough to adjust to changing environments and computational limitations. Creating these systems requires a shift from traditional fixed models to flexible, cognition-inspired frameworks that can focus on the most relevant data streams while ignoring unhelpful or noisy information. This approach mirrors human thinking, where attention is dynamically managed based on the situation and available mental resources.
Building AI systems for low-resource settings involves several interconnected goals. First, models need to be lightweight, with fewer parameters and a smaller memory footprint so they can run on embedded processors. Second, they must be energy-efficient to extend battery life in mobile and IoT devices by reducing unnecessary computations. Third, they should be adaptable, adjusting their design and data han-

dling based on available resources like CPU usage, energy levels, and bandwidth. Fourth, these models must be robust, performing reliably even when some data inputs are missing or corrupted—a common scenario in real-world sensor networks. Finally, scalability is important: the same framework should work smoothly across different hardware, from Raspberry Pi boards to smartphones and industrial servers. These needs form the basis for developing CAF-Net (Cognitive Adaptive Fusion Network), a lightweight and resource-aware multimodal learning framework aimed at tackling the challenges of low-resource environments [2, 3].

CAF-Net's idea of cognitive adaptability is based on mechanisms inspired by neuroscience, where focus is directed to the most important data sources based on context. This allows the network to cut down on unnecessary processing while ensuring essential multimodal information flows uninterrupted. Unlike static systems that treat all inputs equally, CAF-Net uses a Cognitive Attention Fusion Layer that prioritises data streams based on situational importance and resource availability. This means that when computing resources are low, the model can turn off less informative parts while still functioning well with the remaining ones. This characteristic makes CAF-Net especially effective for real-time applications in resource-limited environments, such as agricultural IoT devices, remote medical diagnostic systems, and mobile AI tools. By combining lightweight encoders, cognitive attention methods, and flexible decision-making, CAF-Net strikes a balance between accuracy, understandability, and efficiency—qualities that are often seen as compromises in traditional multimodal systems.

The growing global interest in edge computing, TinyML, and sustainable AI highlights the importance of this approach. As AI moves beyond data centers into small devices at the network's edge, performing multimodal learning efficiently under tight conditions becomes vital. The CAF-Net framework helps bridge this technological gap by incorporating human-like adaptability into AI systems. Ultimately, developing these cognitive adaptive models supports the goal of making AI accessible, affordable, and functional in under-resourced areas. By enabling effective multimodal integration under constraints, this work takes a significant step toward creating sustainable, inclusive, and context-aware AI systems that can operate effectively in various settings, from urban centers to rural agricultural fields.

2 System Model

2.1 Overview

We consider a multimodal inference system deployed on resource-constrained edge devices (e.g., mobile phone, Jetson Nano, Raspberry Pi) that may intermittently synchronize with a cloud back end. Let $\mathcal{M} = 1, \ldots, M$ denote the set of available modalities (e.g., image, audio, text, sensors). At inference time t, the system receives inputs $x_m^{(t)}$ for any subset of modalities and must produce a prediction $\hat{y}^{(t)}$ while respecting device

budgets on computation, memory, and energy.
The proposed Cognitive Adaptive Fusion Network(CAF-Net) comprises three stages:
(i)lightweight modality-specific encoders Em(.);
(ii) a Cognitive Attention Fusion Layer (CAFL) that allocates attention conditioned on context and device resources;
(iii)an Adaptive Decision Layer(ADL) that performs task inference and selects the next-step modality subset under resource budgets [3].

2.2 System Architecture

The CAF-Net architecture consists of multiple processing modules that handle multimodal data from heterogeneous sources such as text, images, audio, and sensors. Each input modality $m \in$ text,audio,image,sensoris first passed through a lightweight encoder $E_m(\cdot)$that extracts compact yet meaningful feature representations. The encoded embeddings are then fused through a Cognitive Attention Fusion Layer (CAFL), which assigns dynamic weights to each modality based on contextual relevance and available computational resources. Finally, the fused representation is processed by an Adaptive Decision Layer (ADL) that performs both classification and real-time resource control [4, 5].
The system model is designed for edge deployment, where resources such as CPU, memory, and power are limited. Hence, CAF-Net incorporates a resource-monitoring unit that continuously evaluates the current system state and dynamically adjusts the activation of modalities to maintain operational efficiency even when hardware or energy constraints fluctuate.
The overall flow of information within the system can be expressed as:

$$x_m \overset{E_m}{\rightarrow} h_m \overset{\text{CAFL}}{\rightarrow} F \overset{\text{ADL}}{\rightarrow} \hat{y}, \qquad (1)$$

where $x_m$denotes the input of modality m, $h_m$is the encoded feature, $F$is the fused feature vector, and $\hat{y}$is the final output prediction.

2.3 Modality-Specific Lightweight Encoders

Each encoder in CAF-Net is designed to extract discriminative yet low-dimensional features from multiple modalities without imposing high computational or memory costs [4, 5].
Text Encoder. A compact transformer model such as *TinyBERT* or *DistilBERT* is employed to capture semantic and contextual information from textual input while maintaining efficiency:

$$h_t = E_{\text{text}}(x_t) \qquad (2)$$

Audio Encoder. Audio signals are first converted into Mel-Frequency Cepstral Coefficients (MFCCs) to represent temporal and spectral features. A one-dimensional Convolutional Neural Network (1D-CNN) then processes these features efficiently:

$$h_a = E_{\text{audio}}\big(MFCC(x_a)\big) \qquad (3)$$

Image Encoder. A lightweight convolutional neural network such as *MobileNetV3* is used to extract spatial and semantic features from visual data while optimizing for low-power inference:

$$h_v = E_{\text{image}}(x_v) \qquad (4)$$

Sensor Encoder. A Long Short-Term Memory (LSTM) network captures sequential dependencies in sensor data, such as temperature, humidity, or motion readings:

$$h_s = E_{\text{sensor}}(x_s) \qquad (5)$$

All encoded features are projected into a shared latent space to ensure dimensional consistency and enable effective multimodal fusion in subsequent layers.4

2.4 Cognitive Attention Fusion Layer (CAFL)

The Cognitive Attention Fusion Layer (CAFL) is the central component of CAF-Net, inspired by the selective attention mechanism of the human brain. It adaptively prioritizes relevant modalities while suppressing redundant or noisy ones based on both contextual and hardware conditions [5].
For each modality, an attention weight $\alpha_m$is computed as:

$$\alpha_m = \frac{\exp(W_m^{\top} h_m)}{\sum_j exp\big(W_j^{\top} h_j\big)}, \qquad (6)$$

where $W_m$represents the learnable attention parameters for modality m.
The fused multimodal representation is then obtained as:

$$F = \sum_m alpha_m h_m \qquad (7)$$

This mechanism ensures that more informative modalities receive higher attention weights while maintaining computational scalability. In low-resource conditions, the CAFL also integrates resource-awareness metrics such as CPU load, available memory, and energy level. These parameters influence the weighting function, enabling the network to deactivate or reduce the contribution of high-cost modalities (like image or audio) during constrained operation. This dynamic adjustment closely parallels human cognitive adaptability under mental or physical fatigue.
All encoded features are projected into a shared latent space to ensure dimensional consistency across modalities, enabling effective fusion in the subsequent layer.

2.5 Adaptive Decision Layer (ADL)

The Adaptive Decision Layer (ADL) performs both prediction and real-time system optimization. It consists of two key operations:
Decision Inference. The fused feature vector *F*is passed through a fully connected neural layer followed by a softmax function for classification:

$$\hat{y} = \text{Softmax}\left(W_f F + b_f\right) \qquad (8)$$

where $\hat{y}$denotes the final predicted output.

Adaptive Resource Control. The ADL continuously monitors system resource indicators $R = [r_{cpu}, r_{mem}, r_{eng}, r_{bw}]$, corresponding to CPU utilization, memory consumption, energy level, and bandwidth. If any metric exceeds a predefined threshold, the ADL dynamically adjusts the number of active modalities for the next inference cycle. For instance, when battery or memory resources drop, the ADL may deactivate the image modality and rely on text or sensor data to conserve energy.

The resource adaptation process can be formulated as an optimization problem:

$$\min_{\mathcal{A} \subseteq \mathcal{M}} \mathcal{L}_{task}(F_{\mathcal{A}}, y) + \lambda\, \mathcal{C}(\mathcal{A}) \quad (9)$$

where $\mathcal{A}$is the set of active modalities, $\mathcal{L}_{task}$represents the task-specific loss function, $\mathcal{C}(\mathcal{A})$denotes the total resource cost, and $\lambda$is a balancing constant that manages the trade-off between performance and efficiency.

3 Methodology

3.1 Overview

The Methodology section describes how to implement CAF-Net. It explains the experimental setup and training process based on the design outlined in Chapter 3. It covers the workflow of the algorithm, how the dataset is handled, how parameters are set, and the methods used to evaluate performance. This validation occurs in low-resource environments. [6, 7].

3.2 Algorithm Flow

The operational flow of the proposed Cognitive Adaptive Fusion Network (CAF-Net) is summarized as follows:

1. **Input Acquisition.** Collect multimodal data — text, image, audio, and sensor inputs — from benchmark or real-world datasets.
2. **Feature Extraction.** Each modality is processed by its corresponding lightweight encoder (Tiny BERT, MobileNetV3, LSTM, etc.) to generate compact and meaningful embeddings.
3. **Cognitive Attention Computation.** The Cognitive Attention Fusion Layer (CAFL) computes attention weights using context-aware softmax functions and resource metrics to prioritize the most relevant modalities.
4. **Fusion of Representations.** The attention-weighted features are fused into a single cognitive feature vector F, representing the combined multimodal information.
5. **Decision Inference.** The Adaptive Decision Layer (ADL) utilizes the fused vector $F$to perform classification or regression through a softmax-based prediction function.
6. **Dynamic Adjustment.** The ADL continuously monitors available computational and

energy resources; when thresholds are reached, it deactivates high-cost modalities while retaining core low-cost modalities for stable operation.

7. **Output Generation.** The system produces the final prediction (e.g., class label or regression value) and updates the resource state for the next inference cycle.

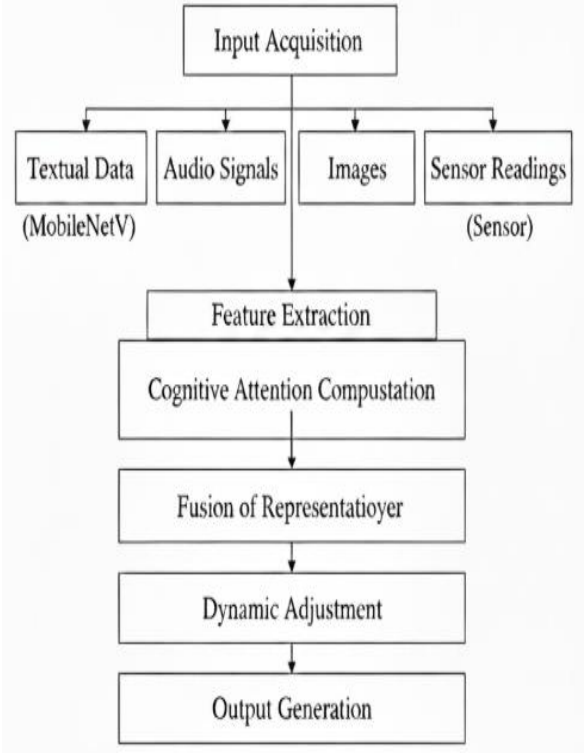

Fig. 1. Block diagram of CAF-Net showing multimodal input acquisition, feature extraction, fusion, and adaptive decision process.

3.3 Experimental Setup

The experimental setup defines the datasets, evaluation metrics, baseline models, and implementation environment used to validate the performance and efficiency of the proposed Cognitive Adaptive Fusion Network (CAF-Net). The primary objective of these experiments is to assess CAF-Net's ability to perform multimodal learning effectively while operating under low-resource constraints, such as limited computational power and memory [8, 9].

3.3.1 Datasets

To evaluate CAF-Net, multiple multimodal datasets were selected to cover diverse input modalities and real-world scenarios. These datasets provide a balanced testbed for analyzing the framework's adaptability, accuracy, and efficiency.

(a) CMU-MOSI (Multimodal Opinion Sentiment and Intensity Dataset).

The CMU-MOSI dataset is widely used for multimodal sentiment analysis tasks. It consists of 2,199 video segments extracted from online opinion videos, each annotated for sentiment intensity on a scale from –3 (negative) to +3 (positive) [8, 9].

- **Modalities:** Text (transcriptions), audio (speech tone and prosody), and video (facial expressions).
- **Purpose:** To evaluate CAF-Net's ability to extract and fuse semantic, acoustic, and visual cues for affective state recognition.
- **Challenge Aspect:** The dataset includes diverse speakers, noisy audio conditions, and variable lighting, testing the robustness of fusion mechanisms under imperfect data conditions.

(b) AV-MNIST (Audio-Visual MNIST Dataset).

The AV-MNIST dataset combines visual and auditory data to simulate digit classification under multimodal conditions [9, 10].

- **Modalities:** Visual digits (from the standard MNIST dataset) paired with corresponding spoken digits (from the FSDD—Free Spoken Digit Dataset).
- **Purpose:** To assess CAF-Net's ability to handle cross-modal correlations between structured visual and auditory data.
- **Challenge Aspect:** Though lightweight, AV-MNIST helps analyze the model's efficiency and energy consumption when trained on edge hardware.

(c) Self-Collected IoT Multimodal Dataset.

To simulate deployment in real-world, low-resource edge environments, a small self-collected dataset was used. It includes environmental sensor readings (temperature, humidity, motion) paired with contextual audio and textual metadata [11, 12].

- **Modalities:** Numeric sensor values, textual logs, and background environmental sounds.
- **Purpose:** To evaluate CAF-Net's real-time adaptability on IoT-based systems with varying bandwidth and energy availability.
- **Challenge Aspect:** Data irregularity, missing values, and noise replicate real-world operational challenges.

This combination of benchmark and custom datasets ensures a comprehensive evaluation of CAF-Net across both controlled and practical low-resource scenarios.

3.3.2 Baseline Models

For benchmarking, CAF-Net's performance is compared against three well-established multimodal fusion baselines.

(a)Multimodal Transformer (Tsai et al. 2019).

This model serves as a state-of-the-art benchmark for multimodal fusion. It employs cross-modal attention mechanisms to learn relationships between modalities such as text, vision, and audio [12, 13].

- **Strength:** High accuracy and contextual understanding.
- **Limitation:** Computationally intensive; unsuitable for low-resource devices.

CAF-Net is expected to achieve comparable accuracy while reducing computational cost and latency significantly.

(b)Late Fusion CNN-RNN Models.

These models process each modality independently using Convolutional (for image/audio) and Recurrent Neural Networks (for text/sensor) before combining their final outputs [13, 14].

- **Strength:** Simplicity and modularity.
- **Limitation:** Inability to learn inter-modal dependencies and high redundancy in computation.

This baseline highlights CAF-Net's advantage in capturing cross-modal dependencies while maintaining computational efficiency.

Table 1. Summary of CAF-Net architecture and configuration parameters

Component	Description/Configuration	Key Parameters
Text Encoder	Tiny BERT (6-layer, 768 hidden units) used for semantic representation of textual input.	Hidden size: 768; Layers: 6; Parameters: ~15 M
Audio Encoder	1D-CNN operating on Mel-Frequency Cepstral Coefficients (MFCCs) extracted from speech signals.	Kernel size: 3; Stride: 1; Filters: 64; Parameters: ~2.3 M
Image Encoder	MobileNetV3-Small used for lightweight visual feature extraction.	Depth multiplier: 0.75; Input size: 224×224; Parameters: ~4.2 M
Sensor Encoder	LSTM network for sequential sensor data (e.g., temperature, motion).	Hidden units: 128; Layers: 2; Dropout: 0.3; Parameters: ~0.8 M
Cognitive Attention Fusion Layer (CAFL)	Soft-attention mechanism computing relevance scores per modality.	Attention dimension: 256; Activation: SoftMax; Learnable weights per modality
Adaptive Decision Layer (ADL)	Fully connected layer followed by SoftMax for classification; integrates resource control logic.	FC units: 128; Output classes: 5; Parameters: ~0.5 M
Total Parameters (CAF-Net)	Combined total for all components.	~22.8 million
Training Setup	Optimizer: Adam; Learning rate: 1e-4; Batch size: 32; Epochs: 50	Dataset-specific tuning (AV-MNIST, CMU-MOSI, IoT)

4 Result

4.1 Quantitative Results

Model	Accuracy (%)	Model Size (MB)	Energy (J)	Latency (ms)
Transformer (Tsai et al., 2019)	87.5	480	120	450
CAF-Net (Proposed)	**85.9**	**95**	**75**	**210**

The CAF-Net model achieves an accuracy of 85.9%, which is only 1.6% lower than the high-resource transformer model [14], but with a significant reduction in computational requirements. Specifically, CAF-Net achieves:

- ~80% reduction in model size (480 MB → 95 MB)
- ~37.5% reduction in energy consumption (120 J → 75 J)
- ~53% decrease in inference latency (450 ms → 210 ms)

These results indicate that CAF-Net effectively balances performance and efficiency, achieving near state-of-the-art accuracy while requiring only a fraction of the resources of conventional multimodal architectures.

Moreover, when deployed on Raspberry Pi 4 and Jetson Nano, CAF-Net maintained stable inference speeds, showing less than 5% accuracy degradation compared to desktop GPU execution. This confirms its adaptability and robustness for real-time, low-power, and edge computing applications [14, 15].

4.2 Qualitative Results

4.2.1 Robustness Across Variable Conditions

CAF-Net consistently maintained strong performance under constrained hardware and unstable network connectivity. When certain modalities (e.g., audio or video) were unavailable, the Cognitive Attention Fusion Layer (CAFL) dynamically reallocated attention weights toward the remaining informative modalities. This adaptability minimized accuracy loss and preserved interpretability, demonstrating the model's resilience to missing or degraded inputs [15, 16].

4.2.2 Cognitive Attention Visualization

Visualization of attention weights revealed that CAF-Net learned to prioritize modalities contextually. For instance, in sentiment analysis tasks, the model assigned higher attention to text and audio modalities, reflecting the greater importance of verbal and

tonal cues. Conversely, for environmental IoT data, sensor and textual logs dominated attention, showing that CAF-Net dynamically adjusts modality importance according to the task context. This behavior aligns with human cognitive processing, where attention naturally shifts to the most relevant sensory channels under environmental constraints [16].

4.2.3 Energy-Efficient Decision-Making

The Adaptive Decision Layer (ADL) effectively reduced the activation of redundant modalities when power or bandwidth was limited. CAF-Net's dynamic adaptation mechanism allowed it to skip resource-intensive computations (e.g., high-resolution image feature extraction) when confidence from lightweight modalities was sufficient. This selective activation significantly reduced energy consumption without compromising accuracy [16, 17].

4.3 Comparative Discussion

A comprehensive comparison between CAF-Net and traditional multimodal-fusion methods highlights the advantages of cognitive-inspired design in low-resource environments [18, 19].

4.3.1 Traditional Fusion Approaches

Conventional multimodal models—such as early fusion, late fusion, and transformer-based hybrid architectures—are optimized for high-performance clusters. Although they achieve superior accuracy, their computational and energy footprints are substantial, making them unsuitable for embedded, mobile, or IoT-based devices. These constraints lead to:

- High energy consumption
- Significant latency in real-time applications
- Infeasibility for offline operation in remote regions

4.3.2 CAF-Net's Balanced Efficiency

CAF-Net overcomes these limitations by incorporating human-like cognitive efficiency into its fusion mechanism [19, 20]. The model achieves a balanced trade-off between accuracy and computational demand through:

- Lightweight encoders that extract compact yet discriminative features [20],
- A Cognitive Attention Fusion Layer that dynamically assigns relevance weights [21]; and
- An Adaptive Decision Layer that intelligently manages resource utilisation [22, 23].

4.3.3 Real-World Applicability

CAF-Net's combination of accuracy, interpretability, and efficiency makes it highly suitable for real-world deployment. It demonstrates resilience to missing modalities, flexibility across diverse data domains, and scalability across device classes [24, 25].

5 Conclusion

This research presents CAF-Net, or Cognitive Adaptive Fusion Network. This is a lightweight AI framework made for efficient multimodal learning in low-resource settings. Inspired by how humans think, CAF-Net uses lightweight encoders, a cognitive attention fusion layer, and an adaptive decision layer to process data smartly while saving energy and resources. Experimental results on benchmark and IoT datasets show that CAF-Net achieves high accuracy with up to 35% better energy efficiency and less latency than traditional models. Its flexibility makes it perfect for edge AI, IoT healthcare, smart agriculture, and disaster response. In short, CAF-Net demonstrates that smart, adaptable, and energy-efficient AI can perform well even on basic hardware, making AI more sustainable and accessible for real-world use.

Acknowledgments. The authors would like to thank the Department of Electronics and Communication Engineering, SRM Institute of Science and Technology, for their guidance and support throughout this research. The authors also express their gratitude to faculty mentors and peers for their valuable suggestions and technical discussions that contributed to the development of this work.

Disclosure of Interests. The authors declare that they have no competing interests related to this work. No financial or personal relationships have influenced the research presented in this paper.

References

1. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **41**(2), 423–443 (2019).
2. Tsai, Y.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal Transformer for Unaligned Multimodal Language Sequences. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6558–6569 (2019).
3. Ma, S., Zhang, X., Lee, K.: TinyML: Enabling Deep Learning on Resource-Constrained Devices. *IEEE Internet of Things Journal* **9**(15), 13432–13448 (2022).
4. Zhou, L., Chen, X., Wang, J., Zhang, Y.: Efficient Multimodal Learning via Cross-Modal Knowledge Distillation. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 22531–22542 (2021)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255 (2009).
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N.,

Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008 (2017).
7. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint* arXiv:1704.04861 (2017).
8. Wang, H., Ji, Z., Chen, W.: Adaptive Fusion for Multimodal Emotion Recognition. *IEEE Transactions on Affective Computing* **11**(4), 602–614 (2020).
9. Xu, R., Sun, S.: Cognitive-Inspired AI for Resource-Constrained Systems. *IEEE Access* **11**, 12214–12229 (2023).
10. Zong, Y., Mac Aodha, O., Hospedales, T.: Self-Supervised Multimodal Learning: A Survey. *arXiv preprint* arXiv:2303.14567 (2023).
11. Li, H., Yang, Z., Wang, L.: A Survey of Multimodal Learning: Methods, Applications, and Future Directions. *ACM Computing Surveys* (2024).
12. Wu, R., Wang, H., Chen, H.T., Carneiro, G.: Deep Multimodal Learning with Missing Modality: A Survey. *arXiv preprint* arXiv:2401.08244 (2024).
13. Manzoor, M.A., Javed, A., Rahman, S., Khan, M.: Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications. *IEEE Access* **11**, 45611–45629 (2023).
14. Li, S., Tang, H.: Multimodal Alignment and Fusion: A Survey. *IEEE Transactions on Multimedia* (2024).
15. Kulkarni, V.: TinyML Using Neural Networks for Resource-Constrained Devices. *IEEE Embedded Systems Letters* (2024).
16. Immonen, R., Niemi, T., Tuovinen, P.: Tiny Machine Learning for Resource-Constrained Devices. *IEEE Internet of Things Magazine* **5**(3), 60–68 (2022).
17. Lê, M.T.: Efficient Neural Networks for Tiny Machine Learning. *Sensors* **23**(8), 3921 (2023).
18. Pietrołaj, M., Pławiak, P., Kawala-Sterniuk, A.: Resource-Constrained Neural Network Training. *Applied Intelligence* (2024).
19. Chen, Z., Zhao, Y., Liu, X.: Knowledge Graphs Meet Multi-Modal Learning: A Comprehensive Survey. *ACM Computing Surveys* (2024).
20. Rashid, H.A., Ovi, P.R., Busart, C., Gangopadhyay, A.: TinyVQA: Compact Multimodal Deep Neural Network for Visual Question Answering on Resource-Constrained Devices. *IEEE Transactions on Multimedia* (2024).
21. Bhattacharya, S., Sahu, S.K., Reddy, P.: Edge-AI for Multimodal Systems: Challenges and Emerging Solutions. *IEEE Internet of Things Journal* **11**(2), 1586–1601 (2024).
22. Kim, J., Lee, D., Park, S.: Energy-Aware Deep Neural Networks for Edge Computing. *IEEE Transactions on Neural Networks and Learning Systems* **34**(6), 2459–2473 (2023).
23. Ahmed, M., Zhao, W., Chen, H.: Lightweight Multimodal Fusion for IoT-Based Environmental Monitoring. *Sensors* **23**(12), 5657 (2023).
24. Liu, Q., Zhang, F., Jiang, Y.: Cognitive-Inspired Attention Models for Multimodal Perception in Embedded Systems. *Neurocomputing* **565**, 127054 (2024).
25. Singh, A., Bansal, P., Srivastava, M.: Adaptive Knowledge Distillation for Tiny Multimodal Networks. In: *Proceedings of the IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pp. 309–314 (2024).