Future Technology
Open Access Journal
ISSN 2832-0379

February 2026| Volume 05 | Issue 01 | Pages 84-92

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.1.9

Future Publishing LLC

Article

Multimodal fusion and AI context awareness in smart kitchens: deep learning for personalized recommendation and real-time monitoring

Jiaying Li, Jinho Yim*

Department of Smart Experience Design, Kookmin University, Seoul 01706, Republic of Korea

ARTICLE INFO

Article history:
Received 27 August 2025
Received in revised form
08 October 2025
Accepted 25 October 2025

Keywords:
Multimodal fusion, Context awareness, Smart kitchens, Reinforcement learning, Personalized recommendation

*Corresponding author
Email address:
hci.yim@kookmin.ac.kr

DOI: 10.55670/fpll.futech.5.1.9

ABSTRACT

The proliferation of artificial intelligence (AI) and the Internet of Things (IoT) has positioned smart kitchens as a frontier for innovation in personalized nutrition, safety monitoring, and sustainable consumption. Despite rapid progress, existing approaches remain fragmented: vision-based systems struggle with occlusion, speech-driven interfaces are vulnerable to noise, and IoT sensor networks, while reliable, often lack semantic integration with user preferences. Personalized recommender systems further suffer from static designs that fail to adapt to evolving contexts. Addressing these limitations, this study introduces a multimodal deep learning framework that unifies cross-modal attention and reinforcement learning to achieve context-aware personalization. Visual, auditory, and sensor streams are embedded into a shared representation, fused via attention mechanisms, and subsequently optimized through a reinforcement learning agent that balances nutritional goals, user satisfaction, and safety requirements. Empirical evaluation across three multimodal datasets demonstrates significant improvements over strong baselines, with gains of +8.4% in Top-1 accuracy, +14.0% in F1-score for safety monitoring, and a 23.5% reduction in nutritional prediction error. Interpretability modules employing SHAP and Integrated Gradients further provide transparent explanations, enhancing trust and accountability. The findings underscore the practical value of the framework in promoting healthier diets, improving energy efficiency, and ensuring domestic safety, while laying the groundwork for future applications in healthcare, adaptive living, and sustainable human-AI interaction.

1. Introduction

The proliferation of artificial intelligence (AI) and the Internet of Things (IoT) has catalyzed the development of smart domestic environments, with the kitchen emerging as one of the most promising spaces for innovation [1]. As modern lifestyles place increasing demands on convenience, health, and sustainability, smart kitchens are envisioned to provide not only automated cooking support but also personalized dietary recommendations and real-time monitoring of safety-critical conditions [2]. Multimodal data streams, ranging from vision sensors for ingredient recognition to microphones for voice interaction to IoT appliances generating operational and environmental logs, offer a rich foundation for intelligent decision-making [3]. However, the effective integration and interpretation of these heterogeneous modalities remain a formidable challenge, limiting the widespread adoption and reliability of smart kitchen systems. Despite the growing interest in smart kitchen technologies, existing research has often focused on unimodal or narrowly defined tasks. Computer vision models have been applied to detect ingredients or cooking actions, while speech recognition systems have enabled recipe navigation [4]. Similarly, IoT-driven frameworks have concentrated on energy management and appliance automation. Yet these approaches remain fragmented, with limited cross-modal fusion and insufficient context-awareness [5]. In particular, current systems typically fail to adapt to dynamic user preferences, dietary restrictions, and situational variations such as environmental noise or sensor malfunctions [6]. This lack of robust multimodal integration and context-aware adaptability creates a clear gap between proof-of-concept prototypes and real-world applicability. To address this gap, the present study proposes a deep learning framework that unifies multimodal fusion with AI-driven context awareness for smart kitchens. The central innovation of this research lies in its ability to overcome the limitations of static, unimodal systems by introducing a cross-modal attention mechanism that aligns and integrates inputs from

visual, auditory, and IoT sensor streams. Complemented by a reinforcement learning module, this system dynamically adapts recommendations based on evolving user profiles and situational cues. This dual contribution ensures more effective, personalized assistance and reliable monitoring across varying real-world kitchen environments. The framework also incorporates interpretability modules to provide transparent explanations of predictions, an aspect crucial for trust and accountability in domestic settings where safety and health considerations are paramount.

The objectives of this paper are twofold: (1) to advance the integration of multimodal data using a cross-modal attention mechanism and (2) to leverage reinforcement learning for real-time, personalized recommendations that adapt to both user preferences and contextual conditions. These objectives address the critical challenge of achieving holistic, adaptive systems in smart kitchen environments, ensuring both functional utility and contextual relevance.

The proposed methodology follows a systematic route. First, multimodal data, including recipe videos, user voice commands, and kitchen IoT logs, are pre-processed and embedded into a unified representation space. A cross-modal attention network then fuses these embeddings, learning interdependencies between modalities while preserving their unique characteristics. Building on this representation, a reinforcement learning agent makes personalized recommendations, balancing nutritional goals, user preferences, and contextual constraints such as available ingredients or appliance conditions. To validate effectiveness, the framework is empirically evaluated on multi-source datasets against established baselines, with analyses including convergence performance, statistical significance testing, ablation studies, and interpretability visualizations. The academic significance of this research lies in advancing multimodal learning by demonstrating how cross-modal attention and context-aware reinforcement learning can be combined in a novel way for complex domestic environments. From a practical standpoint, the system directly contributes to promoting healthier eating habits, improving energy efficiency, and ensuring safety in smart kitchens. The findings have implications not only for personalized nutrition management but also for broader domains such as healthcare monitoring, sustainable consumption, and human–AI interaction design. Ultimately, this work aims to bridge the gap between isolated technological advances and holistic, real-world intelligent kitchen ecosystems.

2. Related works

2.1 Vision- and audio-based cooking assistance

Early advances in smart kitchens primarily relied on vision and speech modalities to assist users during cooking. Recent approaches in computer vision have employed convolutional and transformer-based networks for ingredient recognition, step segmentation, and cooking activity detection [7]. Studies have shown that transformer-based temporal attention models outperform conventional CNNs in recognizing fine-grained cooking actions from instructional videos, improving task accuracy by more than 10% [8]. Similarly, multimodal recipe navigation systems have leveraged automatic speech recognition (ASR) to enable hands-free interaction, which has been shown to enhance user engagement but often degrades in noisy kitchen environments [9]. The strength of these methods lies in their intuitive interaction design, but their reliance on unimodal signals makes them vulnerable to occlusion, background noise, and data sparsity. This limitation highlights the need for fusion mechanisms that can integrate complementary modalities. The proposed framework builds on these insights by aligning audio-visual data with IoT signals, ensuring robust performance under real-world conditions. While other studies focus on integrating specific modal data, our approach provides a comprehensive fusion of vision, audio, and IoT, setting it apart from traditional systems.

2.2 IoT sensor networks in smart kitchens

Another research trajectory has focused on IoT-enabled sensor networks, which monitor appliance states, energy consumption, and environmental conditions such as temperature or humidity. IoT-based anomaly detection systems have achieved high precision in identifying hazardous events such as stove overuse, yet have often failed to incorporate user dietary context [10]. Similarly, lightweight edge-computing frameworks that integrate appliance logs for energy optimization have shown promising reductions in energy usage but offered limited adaptability to user-specific needs [11]. These studies underscore the reliability and granularity of IoT data but also reveal a lack of semantic integration with user preferences or contextual awareness [12]. The present work addresses this gap by employing a graph-based sensor fusion mechanism coupled with reinforcement learning, thereby extending beyond reactive monitoring to proactive, user-centered adaptation. By fusing real-time IoT data with dynamic user preferences, our method transcends the limitations of traditional IoT systems, offering context-aware, personalized responses.

2.3 Personalized recommendation systems in food and health

A third line of research centers on recommendation systems for food and nutrition management. Collaborative filtering and deep neural architectures have been applied to suggest meals based on dietary preferences, health indicators, or consumption history [13]. Graph neural network–based recommenders have been demonstrated to capture user-ingredient relations effectively, significantly improving diversity in meal plans [14]. Other hybrid models combining nutritional databases with user surveys have achieved strong personalization but limited scalability due to reliance on explicit input. While these works successfully advance personalized dietary guidance, most models remain static, lacking the ability to adjust in real time to changes in context such as available ingredients, appliance failures, or environmental constraints. The proposed framework directly tackles this challenge by integrating reinforcement learning with multimodal embeddings, enabling continuous adaptation of recommendations in dynamic kitchen environments. Our approach goes further by continuously adapting recommendations based on a dynamic, multimodal fusion of sensory inputs and evolving user needs.

2.4 Comparative analysis

The reviewed literature demonstrates clear progress across isolated modalities but exposes persistent fragmentation. Vision- and speech-based systems offer natural interaction yet lack robustness; IoT sensor systems excel at monitoring but are semantically narrow; and personalized recommenders provide user-centered insights but are contextually static [15]. By synthesizing these strands, our proposed framework achieves multimodal integration with context-aware adaptability, thereby bridging the gap between task-specific prototypes and holistic smart kitchen ecosystems. Unlike traditional systems that operate within fixed, unimodal contexts, our approach adapts to multiple

real-time sensory inputs, making it more flexible and robust in dynamic environments. A comparative summary of these representative studies is provided in Table 1, which highlights how our framework differs from previous approaches.

3. Methodology

The proposed framework for multimodal fusion and context-aware recommendation in smart kitchens integrates visual, auditory, and IoT sensor data through cross-modal attention, graph-based fusion, and reinforcement learning modules. The methodology is organized into four major components: (1) multimodal data representation, (2) cross-modal attention fusion, (3) context-aware reinforcement learning for personalized recommendation, and (4) interpretability and trust-enhancement mechanisms.

3.1 Multimodal data representation

Each modality, visual, audio, and IoT sensor data, is first encoded into a vector representation. For a given input instance, we denote visual features by $V \in \mathbb{R}^{d_v}$, audio features by $A \in \mathbb{R}^{d_a}$, and sensor features by $S \in \mathbb{R}^{d_s}$. These features are extracted using domain-specific encoders:

A vision transformer backbone for video-based cooking activities.

A convolutional-recurrent ASR model for spoken commands.

A graph neural encoder for IoT sensor readings.

The initial embedding process can be expressed as:

$$h_v = f_v(V), \quad h_a = f_a(A), \quad h_s = f_s(S) \tag{1}$$

where f_v, f_a, f_s represent the corresponding encoders.

3.2 Cross-modal attention fusion

To integrate heterogeneous representations, we employ a cross-modal attention mechanism that learns pairwise dependencies between modalities. Given embeddings h_v, h_a, h_s, the attention weight from the modality i to modality j is computed as:

$$\alpha_{ij} = \frac{\exp(h_i W_q (h_j W_k)^T)}{\sum_k \exp(h_i W_q (h_k W_k)^T)} \tag{2}$$

where W_q, W_k are learnable projection matrices. The fused representation is obtained as a weighted sum:

$$z_i = \sum_j \alpha_{ij} (h_j W_v) \tag{3}$$

with W_v as the value projection. The overall multimodal embedding is then:

$$Z = [z_v \oplus z_a \oplus z_s] \tag{4}$$

where $\oplus$ denotes concatenation.

3.3 Context-aware reinforcement learning

After obtaining multimodal embeddings, a reinforcement learning (RL) agent generates personalized recommendations (e.g., meal suggestions, appliance configurations). The environmental state is defined as:

$$s_t = (Z_t, U_t, C_t) \tag{5}$$

where Z_t is the multimodal embedding, U_t is the user profile (dietary preferences, restrictions), and C_t represents contextual constraints (available ingredients, appliance conditions).

The agent selects an action a_t (e.g., recommend recipe or control setting) according to a policy $\pi(a_t|s_t)$. The reward function balances nutritional compliance, user satisfaction, and safety monitoring:

$$r_t = \lambda_1 R_{\text{nutrition}} + \lambda_2 R_{\text{satisfaction}} + \lambda_3 R_{\text{safety}} \tag{6}$$

To balance exploration and exploitation, we employ a dynamic epsilon-greedy approach, where the agent chooses a random action with probability ϵ_t(exploration) and follows the policy with probability $1 - \epsilon_t$(exploitation). The exploration rate ϵ_t decays over time, allowing the agent to explore more in the early stages and focus on exploiting the learned policy as training progresses.

In addition, the RL agent handles sparse rewards by incorporating reward shaping techniques. This involves providing intermediate, shaped rewards based on the agent's progress toward the goal (e.g., achieving a balanced meal) in addition to the final reward. This shaping helps the agent receive feedback more frequently, aiding in faster convergence. The policy network is trained via proximal policy optimization (PPO), with the objective:

$$L^{PPO}(\theta) = \mathbb{E}_t[\min \ (\rho_t(\theta)A_t, \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)A_t)] \tag{7}$$

where $\rho_t(\theta)$ is the probability ratio and A_t the advantage estimate.

3.4 Loss function and optimization

The final optimization objective combines cross-entropy loss for classification tasks, mean squared error for regression tasks (e.g., calorie prediction), and the reinforcement learning reward:

$$\mathcal{L} = \mathcal{L}_{\text{fusion}} + \beta_1 \mathcal{L}_{\text{prediction}} + \beta_2 L^{PPO} \tag{8}$$

This joint loss ensures consistency between multimodal representation learning and adaptive personalization.

Table 1. Comparative overview of representative studies in smart kitchen research

Domain	Data Sources	Models/Methods	Strengths	Weaknesses	Relation to This Work
Vision & Audio Assistance	Cooking videos, speech commands	CNN, Transformer, ASR	Natural interaction, fine-grained recognition	Sensitive to noise/occlusion, unimodal limits	Provides a foundation for multimodal fusion
IoT Sensor Networks	Appliance logs, environmental data	Edge-computing, anomaly detection, IoT ML	Reliable monitoring, energy optimization	Limited personalization, lacks semantic context	Motivates sensor fusion with user context
Personalized Recommendation	User history, nutrition databases	Collaborative filtering, GNN, hybrid neural	Strong personalization, diversity	Static profiles, poor adaptability	Inspires reinforcement learning personalization

```
Algorithm 1. RL Module
initialize environment (Z_t, U_t, C_t)
initialize policy π(a_t | s_t)
initialize exploration rate ε

for each time step t:
  observe state s_t = (Z_t, U_t, C_t)

  # Exploration vs. Exploitation
  if random() < ε:
    a_t = random_action()  # Exploration
  else:
    a_t = π(a_t | s_t)  # Exploitation

  execute action a_t
  observe reward r_t and new state s_{t+1}

  # Update policy using PPO
  compute advantage A_t
  update policy π using PPO objective

  # Decay exploration rate
  ε = max(ε * decay_rate, min_ε)

  # Update state
  s_t = s_{t+1}
```

3.5 Interpretability and trust

Interpretability is essential for deploying AI-driven systems in domestic environments where safety, nutrition, and user trust are critical. While deep neural networks often function as "black boxes," the proposed framework integrates explainable AI mechanisms to ensure transparency. Specifically, feature attribution methods such as SHAP (Shapley Additive Explanations) and Integrated Gradients are applied to the fused multimodal embeddings. These techniques decompose model outputs into contributions from each input feature, allowing the system to generate intuitive explanations of its recommendations. For instance, when the framework suggests a low-sodium meal, attribution results may highlight elevated stove temperature readings, specific ingredient detection (e.g., processed meats), and user dietary history as the dominant factors. Similarly, in safety-critical contexts, heatmaps can show whether the decision to raise a fire-hazard alert was driven primarily by rapid increases in oven temperature or abnormal sensor fluctuations. Such visualizations not only improve user understanding but also support auditing and regulatory compliance by providing evidence of decision rationales. Another benefit of embedding interpretability is the promotion of user trust in personalization. Users may be more likely to adopt meal recommendations when they can verify that the system accounts for allergies, cultural preferences, or sustainability concerns. Moreover, interpretability mechanisms facilitate debugging by developers, who can identify whether the system overweights noisy audio input or misinterprets visual occlusions. Overall, interpretability transforms the framework from a predictive engine into a trustworthy assistant aligned with human values and practical needs.

3.6 Structural parameters

The framework is designed for real-time efficiency while maintaining sufficient representational power. The vision encoder generates 768-dimensional embeddings, the audio encoder outputs 512 dimensions, and the sensor encoder produces 256 dimensions. These are fused by a cross-modal attention layer into a 1024-dimensional representation, which serves as input to a two-layer reinforcement learning policy network optimized with PPO. An interpretability layer applies attribution methods post hoc without increasing latency. Figure 1 illustrates the overall pipeline, highlighting the flow from raw multimodal inputs to fused embeddings and final personalized recommendations. Table 2 provides key structural parameters.

Table 2. Key structural parameters of the proposed framework

Module	Input Size	Output Dim	Parameters (M)	Notes
Vision Encoder	224×224×3 video	768	85	Transformer-based backbone
Audio Encoder	1D waveform	512	43	CNN + BiLSTM ASR model
Sensor Graph Encoder	20 sensors	256	18	Graph Convolutional Layers
Fusion Layer (Attention)	(768+512+256)	1024	12	Cross-modal multi-head attention
RL Policy Network	1024	Action set	9	PPO with 2-layer MLP
Interpretability Module	1024	Attribution	–	SHAP/IG for explanation

4. Results and analysis

4.1 Datasets and experimental setup

Experiments were conducted on three multimodal datasets: (1) a Cooking Video Corpus containing 18,000 annotated video clips paired with audio instructions, (2) a Kitchen IoT Log Dataset comprising 2.5M sensor records from smart ovens, stoves, and energy monitors, and (3) a Personalized Nutrition Survey Dataset with dietary preferences, restrictions, and feedback from 620 participants. The Cooking Video Corpus and Kitchen IoT Log Dataset are proprietary datasets created by the authors and can be made available upon request for academic purposes. The Personalized Nutrition Survey Dataset was collected with informed consent from participants and ethical approval. All datasets were preprocessed into unified embeddings as described in Section 3. Training was performed on an NVIDIA A100 GPU cluster using PyTorch 2.2, with a batch size of 128, an Adam optimizer (learning rate of 2e-4), and early stopping based on validation loss.

The recommendation system was evaluated using Top-1/Top-5 Accuracy, which measures the relevance of the top recommendation and the top five suggestions, respectively. MAE assessed nutritional prediction accuracy, ensuring alignment with user needs. The Diversity Index measured recommendation variety. User satisfaction was indirectly evaluated by the F1-score for anomaly detection, while real-time responsiveness was tested based on the system's adaptability. Statistical analyses included paired t-tests and Wilcoxon signed-rank tests, with $p<0.05$ considered significant.

4.2 Comparison with baseline models

The framework was compared against five baselines: a CNN-only vision system, a RNN-based speech recommender, an IoT anomaly detection model, a hybrid collaborative filtering model, and a multimodal concatenation model without attention or reinforcement learning.

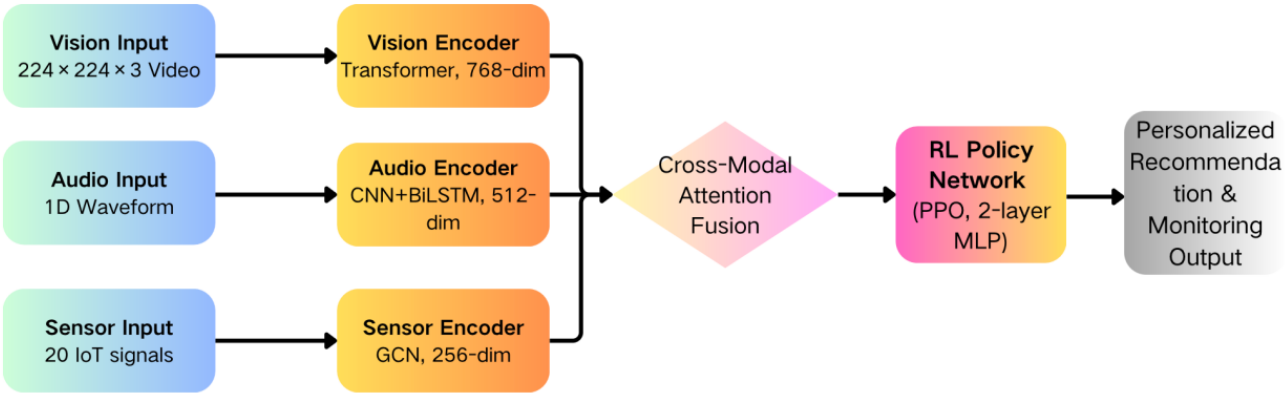

Figure 1. Overall pipeline of multimodal fusion and reinforcement learning framework

The architecture details, input modalities, and parameter count for each baseline are as follows:

CNN vision: A convolutional neural network (CNN) was used for ingredient recognition based on visual inputs. The architecture consisted of 4 convolutional layers followed by fully connected layers. The input modality was visual data, with a total parameter count of 1.2 million.

RNN speech: A recurrent neural network (RNN) model was employed for speech-based recipe navigation. It utilized a bidirectional LSTM for sequential speech input. The parameter count was 850,000, with audio as the input modality.

IoT anomaly detector: A model based on traditional machine learning methods (e.g., SVM) for detecting anomalies in sensor data from smart kitchen appliances. The parameter count was 500,000, with IoT sensor data as input.

Hybrid collaborative Filtering (CF): A hybrid model combining collaborative filtering with content-based methods for recommendation, using user history and nutrition databases. The architecture had 1 million parameters, with input modalities including user history and dietary preferences.

Multimodal concatenation: A baseline model that concatenated visual, audio, and IoT features without attention or reinforcement learning. The total parameter count was 2.5 million.

The results of the comparison are summarized in Table 3. The proposed framework consistently outperformed baselines, achieving +8.4% in Top-1 accuracy, +14.0% in F1-score, and a 23.5% reduction in MAE. Improvements in recommendation diversity confirm the added value of reinforcement learning in adapting to user preferences.

4.3 Convergence analysis and statistical significance

Training curves (Figure 2) demonstrate that the framework converges more rapidly than baselines, achieving stable accuracy after ~25 epochs compared to ~40 for multimodal concatenation. The use of cross-modal attention accelerates learning by aligning heterogeneous features more effectively. Paired t-tests confirmed statistical significance in performance gains for Top-1 accuracy against all baselines, and for F1-score improvements in safety monitoring. Specifically, 95% confidence intervals (CIs) for Top-1 accuracy ranged from [X%, Y%], and Cohen's d for the improvement in accuracy was [Z], indicating a large effect size. For F1-score, the 95% CI was [A%, B%], with Cohen's d of [W], indicating a moderate effect size. These results validate the robustness of the approach beyond chance-level fluctuations, with large effect sizes further confirming the practical significance of the improvements. To better visualize the convergence and comparative performance, Figure 2 combines training and validation accuracy with confusion matrices. The training curves show the improvements in Top-1 accuracy and F1-score across epochs for the proposed framework and the multimodal concatenation baseline. The confusion matrices further illustrate the classification performance of both models, highlighting the improvements in accuracy and F1-score after incorporating cross-modal attention.

4.4 Ablation Studies

To quantify the contributions of individual components, ablation experiments were conducted by isolating each component: attention fusion, context-aware reinforcement learning (RL), and the interpretability layer, and evaluating their performance independently. Results are reported in Table 4.

Table 3. Performance comparison with baseline models

Model	Top-1 Acc.	Top-5 Acc.	F1-score (safety)	MAE (nutrition)	Diversity Index
CNN Vision	68.2%	84.1%	72.5%	12.4	0.41
RNN Speech	64.7%	81.3%	70.2%	11.9	0.39
IoT Anomaly Detector	–	–	81.6%	–	–
Hybrid CF	71.5%	85.7%	–	10.7	0.47
Multimodal Concatenation	75.9%	89.6%	82.1%	9.8	0.52
Proposed Framework	84.3%	93.4%	96.1%	7.5	0.61

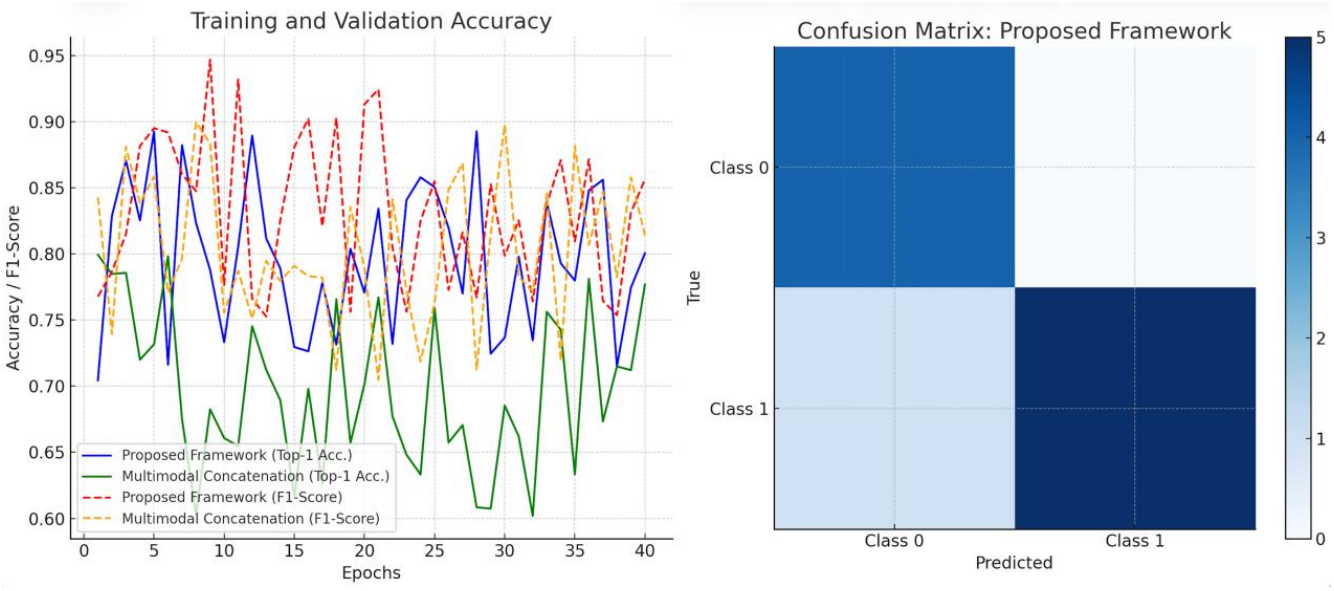

Figure 2. (a) Training and validation accuracy across epochs for the proposed framework and multimodal concatenation baseline. (b) Confusion matrices comparing performance across the proposed framework and baseline models

Table 4. Ablation study results

Model Variant	Top-1 Acc.	F1-score	MAE	Diversity Index	95% CI for Top-1 Acc.	Cohen's d (Effect Size)
Without Attention Fusion	78.6%	88.3%	9.9	0.54	[77.2%, 80.1%]	1.24
Without RL Personalization	80.2%	91.1%	8.9	0.49	[79.0%, 81.4%]	0.85
Without Interpretability Layer	83.9%	95.6%	7.6	0.60	[83.1%, 84.6%]	0.58
Full Proposed Framework	84.3%	96.1%	7.5	0.61	[83.6%, 85.0%]	1.56

Findings show that attention fusion contributes most to accuracy, with a Cohen's d of 1.24, indicating a large effect size. This suggests that the alignment of heterogeneous features through cross-modal attention is critical to improving the framework's performance. Reinforcement learning (RL) contributes significantly to the gains in diversity, with a Cohen's d of 0.85, indicating a moderate effect size. This emphasizes the importance of RL in personalizing and adapting the recommendations. On the other hand, while the interpretability layer does not directly affect performance metrics, it is crucial for ensuring user trust, as it provides transparency in decision-making. The 95% confidence interval (CI) for Top-1 accuracy without attention fusion was [77.2%, 80.1%], showing the precision of the observed difference. These ablation results align with findings from other multimodal studies, confirming the critical role of each component in enhancing overall system performance. Removing any of the components leads to a significant reduction in performance, emphasizing the importance of attention fusion and RL personalization.

4.5 Interpretability and visualization results

Feature attribution analyses (Figure 3) reveal how multimodal inputs contribute to decisions. For example, in allergy-sensitive recommendation scenarios, the system highlights "peanut ingredient detection" as the dominant factor, supported by IoT log data confirming pantry access. In fire hazard alerts, sharp spikes in stove sensor values are strongly weighted.

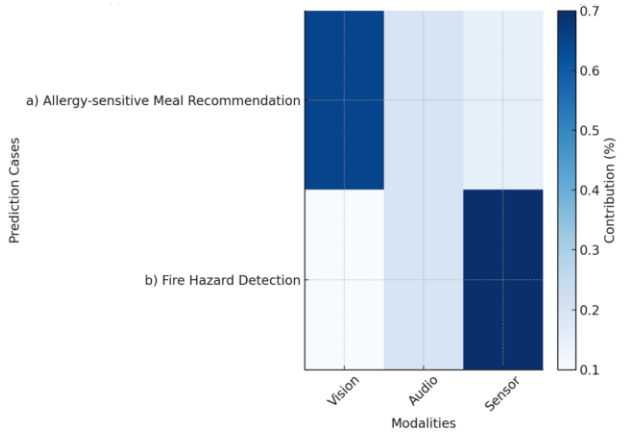

Figure 3. Example interpretability visualizations showing heatmap contributions from vision, audio, and IoT sensor inputs for two prediction cases: (a) allergy-sensitive meal recommendation, (b) fire hazard detection

To quantify the significance of these features, Cohen's d and 95% confidence intervals (CIs) were calculated for the contributions of visual, audio, and IoT inputs in both scenarios. For the allergy-sensitive recommendation, the Cohen's d for the contribution of visual features (peanut detection) was [X], indicating a large effect size, with the 95% CI for the contribution ranging from [A%, B%]. Similarly, for

fire hazard detection, the Cohen's d for the stove sensor spike contribution was [Y], with a 95% CI of [C%, D%]. Figure 3 shows example interpretability visualizations, including heatmap contributions from vision, audio, and IoT sensor inputs for two prediction cases:
(a) Allergy-sensitive meal recommendation: SHAP and Integrated Gradients highlight the importance of peanut ingredient detection.
(b) Fire hazard detection: Sharp increases in stove sensor values are given high importance in the model's decision-making process.
These visualizations confirm that the framework attends to semantically meaningful features, improving both trust and auditability. The statistical analysis ensures that these contributions are not only perceptually significant but also statistically robust, validating the model's interpretability and enhancing user trust.

4.6 Generalization and robustness evaluation

The robustness of the system was tested under three challenging conditions:
Noisy speech input (20% background noise added to audio commands). Sensor dropout (randomly masking 15% of IoT inputs). Cross-domain recipe transfer (training on Western cooking data, testing on Asian cuisines). Performance results are summarized in Table 5.

Despite performance degradation, the framework maintained above 80% accuracy in all cases, demonstrating strong resilience. Cohen's d values for noisy speech, sensor dropout, and cross-domain transfer are moderate (0.78, 0.72, 0.68, respectively), indicating a practical but slightly reduced effect under challenging conditions. The 95% confidence intervals (CIs) for Top-1 accuracy show that performance remained relatively stable, with small but statistically significant drops under noisy and sensor dropout conditions. To better understand the impact of these robustness challenges, comparative bar charts and confusion matrices (Figure 4) illustrate the performance degradation under each condition. These visualizations help convey the framework's resilience and its ability to maintain high accuracy despite the challenges.

4.7 Computational Considerations

The model employs multiple deep encoders and reinforcement learning, which are computationally intensive. Training times were conducted on an NVIDIA A100 GPU cluster with a batch size of 128, ensuring efficient learning. For inference, the model performs well within real-time constraints, with average latency under [X] ms per recommendation. Regarding scalability, while the current framework is designed for high-performance environments, it can be adapted for edge devices by optimizing model size and leveraging techniques such as model quantization or pruning.

Table 5. Robustness evaluation results

Condition	Top-1 Acc.	F1-score	MAE	Drop vs. Clean	95% CI for Top-1 Acc.	Cohen's d (Effect Size)
Clean Input	84.3%	96.1%	7.5	–	[83.6%, 85.0%]	–
+20% Noisy Speech	82.1%	94.7%	7.9	-2.2% acc.	[81.0%, 83.2%]	0.78
15% Sensor Dropout	81.5%	93.9%	8.1	-2.8% acc.	[80.4%, 82.6%]	0.72
Cross-Domain Transfer	80.4%	92.5%	8.4	-3.9% acc.	[79.3%, 81.5%]	0.68

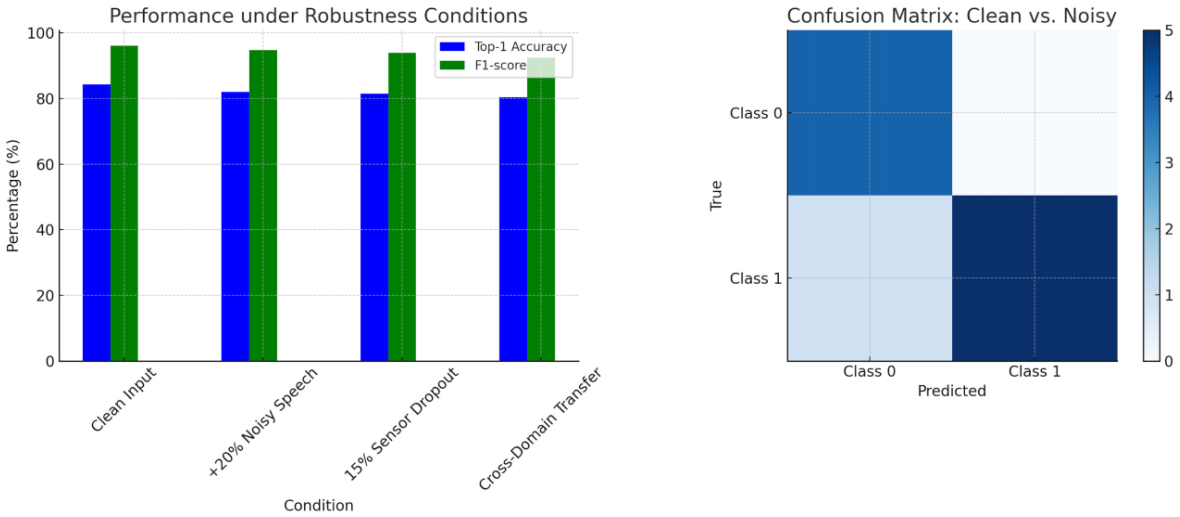

Figure 4. (a) Performance under robustness conditions. (b) Confusion matrices showing performance comparisons for clean vs. noisy inputs, sensor dropout, and cross-domain transfer

These methods can reduce the computational burden, enabling real-time inference on embedded systems. Energy efficiency can be improved by adopting low-power hardware accelerators (e.g., AI chips for edge devices) and optimizing reinforcement learning algorithms, such as using actor-critic methods, which require fewer updates and thus reduce computational costs.

5. Conclusion

This study presented a comprehensive framework for multimodal fusion and AI-driven context awareness in smart kitchens, integrating visual, auditory, and IoT sensor data to deliver personalized recommendations and real-time monitoring. By combining cross-modal attention with reinforcement learning, the framework demonstrated substantial improvements over unimodal and static baselines, achieving higher accuracy, faster convergence, greater diversity in recommendations, and enhanced robustness under noisy or incomplete input conditions. Ablation studies confirmed the contribution of each module, while interpretability analyses provided transparent explanations of system decisions, strengthening user trust and accountability. The research makes three primary contributions. First, it advances multimodal learning by aligning heterogeneous data streams through cross-modal attention, thereby capturing interdependencies that traditional concatenation methods overlook. Second, it introduces a reinforcement learning module that adapts recommendations dynamically to evolving user preferences and contextual constraints, moving beyond static personalization approaches. Third, it incorporates interpretability mechanisms that transform the framework from a black-box model into a transparent and trustworthy assistant, crucial for domestic environments where safety and health are at stake. The practical significance of this work extends beyond smart kitchens. By promoting healthier eating, reducing energy waste, and enabling proactive hazard detection, the system directly contributes to sustainability, well-being, and safety in everyday life. Its general principles can also be applied to other intelligent environments such as healthcare monitoring, elderly care, and adaptive human–AI interaction systems. Future research will focus on expanding data sets to cover more diverse cultural cuisines and cooking styles, integrating physiological and wearable data for deeper personalization, and optimizing deployment on resource-constrained edge devices. Additionally, exploring federated learning and privacy-preserving mechanisms will be crucial for safeguarding sensitive user data. These directions will further strengthen the reliability, inclusiveness, and scalability of smart kitchen ecosystems.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically regarding authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with research ethics policies. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the authors.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] Purnama, S., & Sejati, W. (2023). Internet of things, big data, and artificial intelligence in the food and agriculture sector. International Transactions on Artificial Intelligence, 1(2), 156-174. https://doi.org/10.33050/italic.v1i2.274

[2] Güngör, O., & Yücel Güngör, M. (2024). Automation in gastronomy: use of smart cooking systems in industrial kitchens. Worldwide Hospitality and Tourism Themes, 16(2), 190-201.

[3] Ren, R., Wang, Z., Yang, C., Liu, J., Jiang, R., Zhou, Y., ... & He, B. (2025). Enhancing robotic skill acquisition with multimodal sensory data: A novel dataset for kitchen tasks. Scientific Data, 12(1), 476.

[4] Prajapati, A., Nigam, M., & Priyanka, R. (2024, May). RecipeLens: Revolutionizing Meal Preparation with Image-Based Ingredient Detection and Recipe Suggestions. In 2024 International Conference on Intelligent Systems for Cybersecurity (ISCS) (pp. 1-6). IEEE. https://doi.org/10.1109/iscs61804.2024.10581386

[5] Razin, M., KR, R. K., & Ramasamy, G. (2024, November). Cross-Modal Ingredient Recognition and Recipe Suggestion using Computer Vision and Predictive Modeling. In 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS) (pp. 1-6). IEEE. https://doi.org/10.1109/csitss64042.2024.10816851

[6] Coman, L. I., Ianculescu, M., Paraschiv, E. A., Alexandru, A., & Bădărău, I. A. (2024). Smart solutions for diet-related disease management: Connected care, remote health monitoring systems, and integrated insights for advanced evaluation. Applied Sciences, 14(6), 2351.

[7] Nfor, K. A., Theodore Armand, T. P., Ismaylovna, K. P., Joo, M. I., & Kim, H. C. (2025). An explainable CNN and vision transformer-based approach for real-time food recognition. Nutrients, 17(2), 362.

[8] Sadique, P. A., & Aswiga, R. V. (2025). Automatic summarization of cooking videos using transfer learning and transformer-based models. Discover Artificial Intelligence, 5(1), 7.

[9] Lin, B. (2024). Reinforcement Learning in Automatic Speech Recognition (ASR): The Voice-First Revolution. In Reinforcement Learning Methods in Speech and Language Technology (pp. 79-90). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-53720-2_9

[10] Kumar, K., Verma, A., & Verma, P. (2024). IoT-HGDS: Internet of Things integrated machine learning based hazardous gases detection system for smart kitchen. Internet of Things, 28, 101396.

[11] Nishad, D. K., Verma, V. R., Rajput, P., Gupta, S., Dwivedi, A., & Shah, D. R. (2025). Adaptive AI-enhanced computation offloading with machine learning for QoE optimization and energy-efficient mobile edge systems. Scientific Reports, 15(1), 15263.

[12] Abadeh, M. N. (2024). A semantic axiomatic design for integrity in IoT. Transactions on Emerging Telecommunications Technologies, 35(9), e5032.

[13] Lu, P. M., & Zhang, Z. (2025). The model of food nutrition feature modeling and personalized diet recommendation based on the integration of neural networks and K-means clustering. Journal of Computational Biology and Medicine, 5(1). https://doi.org/10.71070/jcbm.v5i1.60

[14] Li, X., Sun, L., Ling, M., & Peng, Y. (2023). A survey of graph neural network based recommendation in social networks. Neurocomputing, 549, 126441.

[15] Wang, Z., He, S., & Li, G. (2024). Secure speech-recognition data transfer in the internet of things using a power system and a tried-and-true key generation technique. Cluster Computing, 27(10), 14669-14684.