DietWatch: Fine-grained and Robust Dietary Monitoring via Smartwatch in Real-World Scenarios

Zhen Hou, Yucheng Xie, Feng Li, *Member, IEEE,* Chengyi Liu, and Honggang Wang, *Fellow, IEEE*

***Abstract*—Dietary behaviors play a pivotal role in promoting overall health and preventing chronic diseases (e.g., hypertension and diabetes). The widespread adoption of smartwatches offers a promising platform for continuous dietary monitoring. However, existing smartwatch-based dietary monitoring approaches struggle with challenges in real-world scenarios, including dynamic interference, gesture generalization, and user diversity. To address these limitations, we propose *DietWatch*, a real-world dietary monitoring system that utilizes a commercial smartwatch to capture and analyze fine-grained dietary behaviors. *DietWatch* incorporates a dynamic interference mitigation module to suppress acoustic and inertial noise. It further employs a contrastive learning-based framework to distinguish eating gestures from diverse daily activities, without constraining users' eating styles and activity types. To enhance generalizability across users, *DietWatch* adopts a cross-user adaptation mechanism to extract user-independent features. Furthermore, a clustering algorithm is designed to estimate dietary time, while an attention-based multimodal fusion method is employed to analyze biting and chewing frequencies and identify food categories. Experimental results demonstrate that *DietWatch* achieves 79.95% temporal Intersection over Union for eating time detection, 85.68% accuracy in food classification, and mean absolute errors of 1.26 bites/min for biting frequency and 7.71 chews/min for chewing frequency estimation.**

***Index Terms*—Dietary monitoring, smartwatches, Internet of Things (IoT), multimodal sensor fusion, contrastive learning, cross-user adaptation.**

I. Introduction

DIETARY behavior has a profound impact on human health. According to reports from the World Health Organization, poor dietary habits are key risk factors for chronic conditions, including obesity, diabetes, and cardiovascular diseases [1]. Given the significant health risks associated with poor diet, it becomes essential to monitor and understand individual dietary behaviors—including dietary time, biting and chewing frequency, and food category selections. These detailed insights are key to uncovering underlying dietary issues, providing a basis for developing effective health intervention strategies, and delivering personalized nutritional advice. For instance, identifying the timing of dietary intake could reveal snacking habits, which in turn can uncover hidden contributors to weight gain [2]. Tracking biting and chewing frequency can reveal behaviors like rapid consumption, which often leads to overeating and poor digestion [3]. In addition, understanding food categories consumed is crucial for nutritionists to create balanced and personalized dietary plans tailored to individual needs [4].

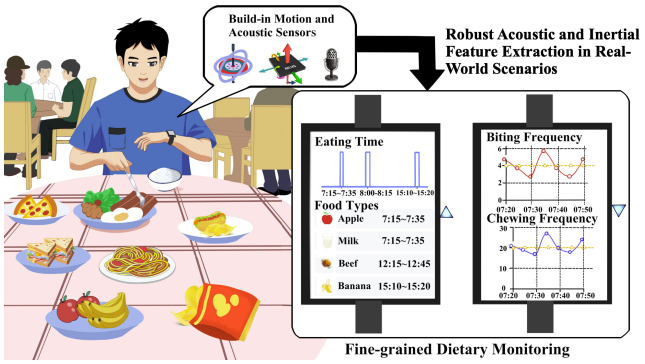


Fig. 1. Illustration of our fine-grained and robust dietary monitoring via smartwatch. DietWatch captures and analyzes eating behaviors in real-world scenarios.

Traditional dietary behavior analysis approaches rely on self-reporting tools, including food diaries and 24-hour recall questionnaires [5]. While commonly employed, these methods suffer from subjective bias and memory errors, compromising the accuracy of dietary information on meal timing and food types. Moreover, they cannot monitor biting and chewing frequency, limiting their ability to provide a complete picture of eating behaviors [6]. To overcome these limitations, researchers propose employing wearable devices for dietary monitoring [7]–[11], which are worn on the body to facilitate continuous dietary-related data collection and enable automated monitoring. Specifically, researchers utilize head-mounted cameras [12], [13] to capture videos associated with the texture of foods. Customized inertial sensors worn on the upper limb capture food-grab gestures, providing insights into hand-to-mouth actions during eating [14], [15]. Dedicated devices such as digital laryngographs worn around the neck and RF tri-layer sensors mounted on teeth detect chewing behaviors [16], [17]. Despite demonstrating high accuracy, these wearable devices face significant practical limitations. They are often cumbersome, making users self-conscious in social settings, particularly during public dining. Additionally, high costs and operational complexity limit their accessibility for general users.

Due to the widespread adoption and relatively low cost,

Manuscript received Month DD, 2025; revised Month DD, 2025; accepted December 03, 2025. *(Corresponding author: Yucheng Xie.)*

Zhen Hou is with the Luddy School of Informatics, Computing, and Engineering, Indiana University Indianapolis, Indianapolis, IN 46202 USA (e-mail: houzhen@iu.edu).

Yucheng Xie, Chengyi Liu, and Honggang Wang are with the Department of Graduate Computer Science and Engineering, Katz School of Science and Health, Yeshiva University, New York, NY 10016 USA (e-mail: yucheng.xie@yu.edu; cliu7@mail.yu.edu; honggang.wang@yu.edu).

Feng Li is with the Department of Computer and Information Technology, Purdue University in Indianapolis, Indianapolis, IN 46202 USA (e-mail: fengli@purdue.edu).

smartwatches have emerged as an ideal platform for daily dietary monitoring. With over 224 million users worldwide and prices starting as low as tens of dollars, smartwatches provide an accessible and cost-effective way for continuous health tracking [18], [19]. Moreover, smartwatches integrate multiple sensors, such as inertial measurement units (IMUs) and microphones, allowing for comprehensive dietary feature extraction from inertial and acoustic domains and enabling fine-grained monitoring. Existing smartwatch-based eating monitoring methods have shown initial success in tracking various aspects of dietary behaviors [15], [20]–[24]. However, these approaches remain limited in scope, providing only partial tracking of dietary behavior, such as eating time. To gain a comprehensive understanding of dietary habits, there is a significant need for a fine-grained monitoring system that captures multiple dietary behaviors simultaneously, including eating time, biting/chewing frequency, and food category selections. A more pressing limitation is that most existing smartwatch-based eating monitoring methods only perform optimally in controlled conditions. The diverse and unpredictable nature of real-world scenarios necessitates a more practical solution.

To bridge the gap, several key challenges need to be addressed, including dynamic interference, gesture variability, and user diversity. *(1) Dynamic Interference.* Dynamic interferences, including motion-induced artifacts (e.g., repositioning utensils or walking) and background noise (e.g., music in public environments) are unpredictable. They can significantly affect data collected from the smartwatch, leading to errors in dietary behavior recognition. Mitigating the impact of such interference is essential for enhancing the robustness of dietary monitoring. *(2) Gesture Generalization.* Consistent features of intake gestures must be extracted, as motion range and speed can vary even when the same gesture is performed across different scenarios or emotional states. Furthermore, requiring users to provide samples for every potential non-eating activity during training is impractical. An effective real-world dietary monitoring system must accurately identify eating times even when encountering previously unseen daily activities. *(3) User Diversity.* Individual differences in body shape and oral structure present challenges for systems trained on limited subject data, as they may struggle to generalize to unseen users without retraining on user-specific data. Therefore, it is necessary to develop a system that can accommodate unseen users without extra dietary sample collection. In addition to addressing these challenges, the system should provide fine-grained and comprehensive monitoring in real-world settings. This includes dietary time recognition, chewing and biting frequency estimation, and food category classification. Such a system would enable accurate dietary behavior tracking and support future efforts in dietary assessment and personalized nutrition.

To address these challenges, we propose a real-world dietary monitoring system *DietWatch*, which achieves fine-grained and robust dietary behavior monitoring using a commercial smartwatch as shown in Figure 1. (1) We develop an interference mitigation method for efficiently reducing real-world interference in both acoustic and inertial domains. Specifically, we develop a self-supervised denoising module that integrates a Bidirectional Gated Recurrent Unit (Bi-GRU) encoder and a Least Mean Square (LMS) adaptive filter to suppress motion-induced inertial noise. We also develop a time-domain convolutional neural network (Conv-TasNet) [25] to build adaptive masks for the accurate extraction of dietary-related acoustic signals. (2) To extract consistent features of eating posture and enable robust recognition, we develop a contrastive learning-based eating gesture identification method, which enhances discriminability of temporal features between eating and non-eating activities through contrastive feature alignment. (3) To enhance the robustness of *DietWatch* to individual differences in dietary behaviors, we develop a cross-user adaptation module using an adversarial autoencoder with maximum mean discrepancy regularization. By extracting dietary behavior features impervious to individual variations, *DietWatch* can function effectively with unseen individuals without extra training efforts. To achieve fine-grained dietary monitoring, we design a clustering algorithm to estimate dietary time based on eating gesture identification. We also develop an attention-based multimodal fusion strategy that combines inertial and acoustic features, enabling fine-grained estimation of biting and chewing frequencies, as well as food category classification.

This paper is an extension of our preliminary work presented at IEEE/ACM CHASE 2025 [26], with additional modules and comprehensive real-world evaluations. Our main contributions are summarized as follows:

- We propose *DietWatch*, a robust dietary monitoring system using a commercial smartwatch in real-world scenarios, providing fine-grained dietary information including dietary time, biting and chewing frequencies, and food categories.
- We develop an interference mitigation approach using Bi-GRU and LMS filters to reduce inertial noise and Conv-TasNet to isolate eating-related acoustic signals. We further develop a contrastive learning-based method to recognize dietary gestures without restricting on users' eating style and daily activities.
- We design a multimodal feature fusion strategy with attention mechanism to effectively combine dietary features from different domains, and develop a cross-user adaptation module that mitigates user-specific dietary behavior differences, reducing the training efforts for unseen users.
- We evaluate *DietWatch* on 40 food types in real-world scenarios. Experimental results demonstrate that *DietWatch* achieves robust performance with 79.95% temporal Intersection over Union (tIoU) for eating time detection, 85.68% accuracy for food type classification, and mean absolute errors of 1.26 bites/min for biting frequency and 7.71 chews/min for chewing frequency estimation.

II. Related Work

A. Dietary Monitoring Using Customized Wearable Sensors

In order to monitor dietary behaviors, researchers have explored using sensors attached to different parts of the body

TABLE I
FUNCTIONAL COMPARISON BETWEEN DIETWATCH AND EXISTING SMARTWATCH-BASED DIETARY MONITORING APPROACHES.

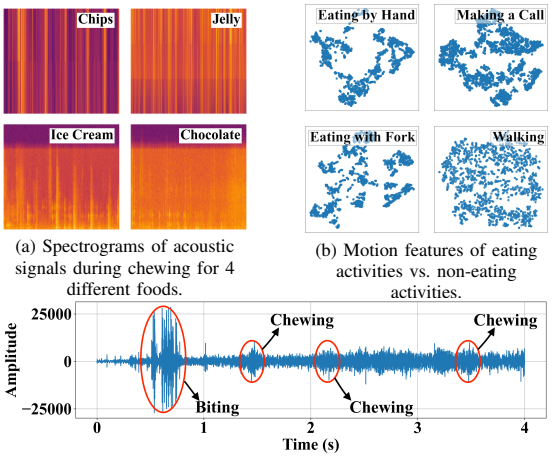
Dietary Monitoring Systems	Real world	Eating time	Food types	Bite/Chew frequency
Sharma et al. [15]	×	✓	×	×
Kyritsis et al. [34]	×	✓	×	✓
Zhang et al. [31]	×	×	×	✓
Thomaz et al. [35]	×	×	×	×
Wang et al. [20]	×	✓	×	✓
Sen et al. [36]	×	✓	✓	✓
Zhang et al. [30]	×	✓	✓	×
DietWatch	✓	✓	✓	✓

to capture the sound, motion, images, or physiological signals associated with eating behaviors. For instance, wearable microphone sensors, such as ear-worn earpieces and throat-mounted microphones, monitor food categories by capturing chewing and swallowing [7], [8], [11], [27]. Smart cameras, including intelligent glasses or head-mounted devices, monitor dietary intake by capturing images of food [9], [13]. IMU sensors embedded in accessories like eyeglasses, wristbands, and necklaces identify food intake or chewing behaviors by detecting characteristic vibrations [28]–[31]. More specialized devices, such as neck-worn EGG sensors, head-mounted EMG sensors, arm-mounted microneedles, and tooth-mounted RF-trilayer sensors, have been explored to capture physiological signals for dietary monitoring [16], [17], [32], [33].

Although these technologies demonstrate high accuracy in detecting eating behaviors, they have several limitations that reduce their practicality for everyday use. Many sensors require precise placement or specialized configurations, causing discomfort during prolonged use and limiting user adherence, particularly for children and the elderly. Additionally, the need for professional setup and the high cost of these devices further restrict their accessibility and large-scale deployment.

B. Dietary Monitoring Using Smartwatches

Smartwatches are widely adopted commercial devices, with approximately 224 million users globally as of 2024 [18], [19]. Their high acceptance rates and integration of multiple sensors make them particularly suitable for real-world dietary monitoring. Internet of Things (IoT) technologies have also been explored for wearable health monitoring. Huang et al. integrated EMG sensors into smart glasses for dietary monitoring, transmitting chewing activity data to cloud platforms via IoT frameworks [37]. Broader IoT healthcare applications have demonstrated the integration of wearable sensors and body area networks for real-time health monitoring [38]. Recent work has applied IoT paradigms to wrist-worn sensor nodes for physiological monitoring [39]. However, these IoT-based approaches typically focus on single-parameter tracking rather than comprehensive dietary behavior analysis in open-world scenarios. Researchers have utilized smartwatches for dietary time estimation by recognizing eating gestures [15], [21], [23], [31], [36]. However, these methods often perform poorly in real-world scenarios due to the presence of previously unseen non-eating activities, such as touching the face or brushing



(a) Spectrograms of acoustic signals during chewing for 4 different foods.

(b) Motion features of eating activities vs. non-eating activities.

(c) Acoustic waveform of biting and chewing in a restaurant.

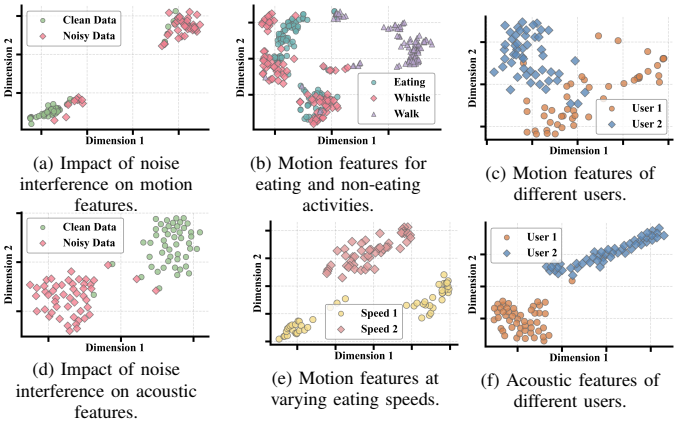
Fig. 2. Characteristic acoustic and motion signals captured by smartwatches for eating behavior detection.

teeth, which can be misclassified as eating gestures. Smartwatches have also been employed for food category classification [36], [40], but existing studies are limited by small food category sets (fewer than 7 types) and insufficient handling of dynamic interference and user variations. Similarly, biting and chewing frequency estimation methods face challenges due to oral structure variations across users, making them difficult to generalize without additional data collection [8], [30], [36], [40], [41]. To the best of our knowledge, existing systems typically provide partial tracking of dietary behaviors, focusing on a single aspect such as dietary time or food type. In contrast, our system offers fine-grained dietary monitoring and is designed to perform robustly in real-world scenarios. A detailed comparison between our system and existing dietary monitoring approaches is provided in Table I.

III. PRELIMINARIES

A. Feasibility Study

Smartwatches can monitor eating behaviors using motion and acoustic signals. Wrist movements, such as grabbing food and bringing it to the mouth, generate distinct inertial data captured as sequences of acceleration and angular velocity. Simultaneously, biting and chewing produce acoustic signals whose characteristics vary with food texture due to differences in bone conduction and air propagation. To demonstrate the feasibility of using smartwatches for fine-grained eating behavior monitoring, we analyzed smartwatch data from multiple participants consuming various foods. Figure 2(a) illustrates spectrograms of chewing sounds from foods such as chips, jelly, ice cream, and chocolate, revealing unique frequency patterns associated with their textures. Figure 2(b) presents a t-SNE visualization [42] of motion features, distinguishing between eating gestures (e.g., eating with forks) and non-eating activities (e.g., making a call). Additionally, Figure 2(c) shows acoustic waveforms collected during eating, where high-amplitude regions (marked in red) represent individual biting or chewing events.



(a) Impact of noise interference on motion features.
(b) Motion features for eating and non-eating activities.
(c) Motion features of different users.
(d) Impact of noise interference on acoustic features.
(e) Motion features at varying eating speeds.
(f) Acoustic features of different users.

Fig. 3. Key challenges of eating behavior monitoring in real-world scenarios.

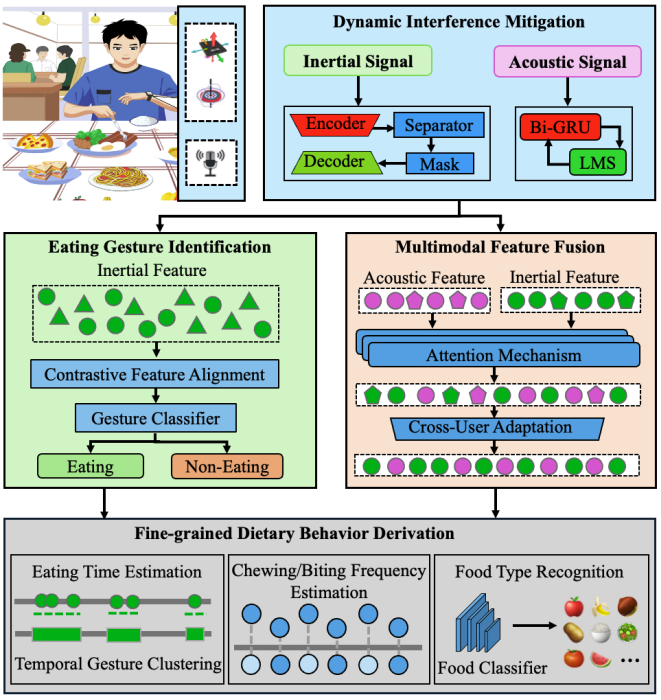


Fig. 4. DietWatch system design.

B. Key Challenges in Real-World Dietary Monitoring

Developing a practical dietary monitoring system that can be used in real-world scenarios faces several key challenges:

Dynamic Interference. Dynamic interference poses significant challenges to both inertial and acoustic data collection in real-world scenarios [7], [43]. For inertial data, considerable noise may arise from unstructured hand motions (e.g., repositioning utensils, adjusting plates) or environmental vibrations caused by walking or riding on public transportation. These motion-induced artifacts, as illustrated in Figure 3(a), distort speed and acceleration measurements, compromising the distribution of motion features extracted from eating-related inertial signals. For acoustic data, interference sources such as conversations, keyboard typing, TV, and traffic noise often overlap with the frequency bands of eating behaviors such as chewing and biting sounds [40]. As shown in Figure 3(d), these noises distort critical acoustic features of eating behaviors.

Gesture Generalization. Recognizing eating gestures is challenging in real-world scenarios due to the prevalence of hand-to-face movements resembling eating gestures, such as applying makeup, smoking, and adjusting glasses [43], [44]. As shown in Figure 3(b), these non-eating activities often generate motion features that overlap with eating gestures, leading to false detections. One potential solution is to collect a wide range of non-eating activities to better delineate the decision boundaries separating eating from non-eating activities in feature space [35]. However, collecting training data for every possible non-eating activities is impractical due to the vast diversity of daily activities. This challenge is further exacerbated by variability in eating behavior across different scenarios, emotional states, and individual habits. For instance, the speed and range of eating gestures may differ significantly between casual dining at home and formal restaurant settings or between hurried weekday meals and relaxed weekend dining [14]. Figure 3(e) illustrates the motion features of eating gestures performed at varying speeds. These challenges necessitate the development of a more robust method for eating behavior detection.

User Diversity. Individual differences in eating behaviors pose a major challenge to system generalization [43]. Our preliminary results reveal substantial variations in inertial and acoustic features among users, even when performing identical eating behaviors. For example, as shown in Figure 3(c) and (f), different users exhibit distinct motion and acoustic feature patterns while consuming the same foods. These individual differences pose a challenge to developing a universal eating monitoring system that generalizes across users. Existing methods typically adopt one of two approaches: retraining on unseen user data, or pre-training on a large, diverse dataset to build a robust model [14], [34], [35]. However, these strategies are often impractical in real-world scenarios due to the high burden of data collection, as users may be unwilling or unable to provide the extensive training data required [40].

IV. DietWatch Design

To address these challenges, we develop *DietWatch*, a fine-grained dietary monitoring system design for real-world scenarios with minimal user effort. The core of *DietWatch* consists of four components: dynamic interference mitigation, contrastive learning-based eating gesture identification, multimodal feature fusion with cross-user adaptation, and fine-grained dietary behavior derivation as shown in Figure 4.

DietWatch takes synchronized time-series inertial and acoustic readings from smartwatches as input. To mitigate noise in both modalities, we design a dynamic interference mitigation module. In the inertial domain, we develop a self-supervised denoising module that integrates a Bi-GRU encoder and a Least Mean Square (LMS) adaptive filter to suppress motion-induced noise. In the acoustic domain, we adopt Conv-TasNet to isolate eating-related sounds by generating adaptive time-frequency masks. To accurately recognize eating gestures among diverse daily activities, we develop a GRU-based feature extractor that captures temporal features from

denoised sequential inertial data and applies contrastive feature alignment to enhance discriminability between eating gestures and other daily activities. Building on this foundation, we design a multimodal feature fusion module with an attention mechanism, enabling the system to emphasize eating-relevant signals. We also introduce a cross-user adaptation strategy to remove user-specific biases via adversarial learning, facilitating generalization to unseen users without retraining.

V. Dynamic Interference Mitigation

A. Inertial Signal Processing

In real-world scenarios, inertial signals collected from smartwatches are inevitably affected by motion-induced noise, including unstructured hand motions (e.g., repositioning utensils) or environment-induced vibrations (e.g., those generated by traveling in a moving vehicle). To address this challenge, we develop a self-supervised denoising approach based on Bi-GRU [45] and Least Mean Square (LMS) [46] adaptive filter, which does not require explicit noise annotations during training. We choose Bi-GRU because eating gestures exhibit distinct temporal patterns, such as periodic wrist motions, which Bi-GRU can effectively capture by leveraging its ability to process sequential data in both forward and backward directions. In contrast, unstructured hand motions (e.g., repositioning utensils, adjusting plates, or switching hand positions) lack temporal coherence, often appearing as abrupt, irregular signal deviations. During self-supervised training with our reconstruction and smoothness objectives, the Bi-GRU learns to preserve the structured periodicity of eating gestures while reducing sensitivity to unstructured noise signals [47].

To further reduce the impact of environment-induced vibrations, we employ a LMS, which dynamically attenuates environment-induced noise by iteratively adjusting to signal variations in real-time. The algorithm minimizes the mean squared error (MSE) between noisy and reference signals by iteratively updating its filter weights. Reference signals are pre-collected in controlled conditions (e.g., eating gestures recorded in low-noise environments). To improve robustness, we incorporate a confidence-based online refinement mechanism that updates reference signal statistics only when both low environmental noise and high-confidence eating gestures are identified [46].

Self-supervised Multi-objective Denoising Strategy. We design a multi-objective self-supervised loss function to denoise signals without requiring labeled data. The training objective consists of a reconstruction loss and a temporal consistency loss, complemented by two regularization terms. The overall loss is:

$$L = \lambda_1\|\hat{\mathbf{x}} - \mathbf{x}\|^2 + \lambda_2 \sum_{t=1}^{T-1} \|\hat{\mathbf{x}}_{t+1} - \hat{\mathbf{x}}_t\|^2 + \lambda_3\|\hat{\mathbf{x}} - \mathbf{r}\|^2 + \lambda_4\|\hat{\mathbf{x}} - \mathbf{x}_{LMS}\|^2, \quad (1)$$

where $\hat{\mathbf{x}}$ is the reconstructed signal generated by the Bi-GRU model, $\mathbf{x}$ is the raw input signal, $\mathbf{r}$ is the representative gesture template (e.g., average of multiple samples collected under low-noise conditions), and $\mathbf{x}_{LMS}$ is the denoised output of an adaptive LMS filter.

Each term in the loss function plays a complementary role: The first term enforces fidelity to the original signal. The second encourages temporal smoothness to suppress high-frequency noise. The third aligns the output with the pre-collected reference signals. The fourth regularizes the output toward an independently filtered version from LMS. The weights λ_1-λ_4 are tuned empirically to balance reconstruction accuracy and denoising effectiveness.

B. Acoustic Signal Processing

In addition to motion-induced noise in inertial signals, acoustic signals captured during eating also suffer from significant interference, especially in real-world scenarios. Analyzing eating-related sounds is crucial, as these signals provide information about food texture and the presence of chewing/biting. However, acquiring high-quality eating sounds in real-world scenarios is challenging due to environmental noise, particularly in public spaces where noise levels vary unpredictably in intensity (50–70 dB) and frequency (100 Hz-10 kHz). Traditional frequency-domain denoising methods, such as Wiener filtering and spectral subtraction, are inadequate as they often distort the signal and fail to preserve subtle acoustic features critical for differentiating food textures [48].

Multi-branch Conv-TasNet Design. To overcome these limitations, we design a time-domain denoising approach based on a multi-branch Conv-TasNet architecture [25]. As shown in Figure 5, we develop an encoder that transforms input waveforms into high-dimensional feature representations using convolutional filter banks. An attention module is trained jointly with the rest of the network to compares these features with texture representations learned by each branch and assigns weights to each Temporal Convolutional Network (TCN) branch accordingly [49]. Each branch is specialized in a particular food texture category and generates a separation mask to enhance texture-relevant acoustic components while suppressing irrelevant noise.

Training and Denoising Procedure. The Conv-TasNet is trained on a curated dataset [25], augmented with environmental noises via an SNR-based mixing method [50]. To promote specialization, the dataset is grouped into food texture categories (e.g., crispy, soft, mixed), and each TCN branch is trained predominantly on one category using a weighted loss:

$$L = \sum_{i=1}^{K} \alpha_i \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2 + \lambda \sum_{i=1}^{K} \|\mathbf{M}_i\|_1, \quad (2)$$

where $\hat{x}_i$ and x_i represent the reconstructed and clean signals for the i-th texture category, M_i is the separation mask, α_i is a texture-specific weight, and λ controls the sparsity of the mask. During inference, the attention mechanism assigns a weight to each branch by comparing input features with the texture representations learned by each branch. The final denoised signal is reconstructed by aggregating the outputs of all branches, weighted according to these attention scores.

VI. Contrastive Learning-based Eating Gesture Recognition

Eating gestures are distinct motion patterns involved in food consumption, including utensil manipulation, food preparation

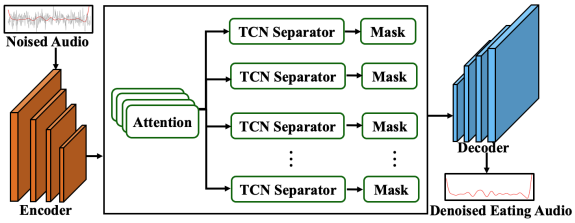


Fig. 5. The multi-branch Conv-TasNet architecture for acoustic signal denoising.

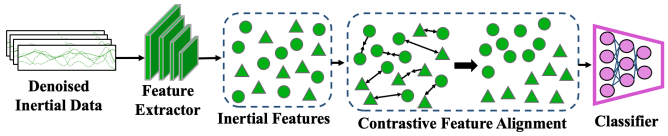


Fig. 6. The contrastive learning-based framework for eating gesture identification.

(e.g., stirring, cutting), and hand-to-mouth movements. [43], [51]. Recognizing eating gestures in real-world scenarios presents unique challenges due to unseen daily behaviors and the variability of unrestricted gestures (e.g., variations caused by individual habits, emotional states, and eating scenarios) as outlined in Section III-B. Traditional supervised learning methods, which rely on predefined gesture categories and extensive training data, often fail to generalize in real-world scenarios due to the labor-intensive task of collecting exhaustive samples for every possible non-eating activity and the variability of eating gestures [35].

Contrastive Learning Framework. To address this issue, we propose a contrastive learning framework that extracts robust, invariant features capturing the shared characteristics of diverse eating gestures, while distinguishing them from kinematically similar non-eating activities [43]. The core idea is to exploit the temporal dynamics (e.g., motion periodicity and consistency) and spatial characteristics (e.g., range and directionality) of gesture. These features capture similarities across eating gestures and differences when contrasted with non-eating activities. By learning invariant features over diverse activities, the framework enhances generalization to previously unseen or unrestricted activities. As illustrated in Figure 6, the proposed framework consists of three main components: feature extraction, contrastive feature alignment, and classification. The feature extraction module processes inertial data to derive representative embeddings, which are further refined by contrastive alignment to enhance their discriminative power. The aligned embeddings are then fed into a classifier to enable robust discrimination between eating and non-eating activities, supporting generalization to diverse and unrestricted eating behaviors.

Loss Function Design and Optimization. We design a contrastive loss function specifically for distinguishing eating gestures from similar non-eating gestures while accounting for variations in eating styles. The feature extractor $f(\cdot)$ is trained to minimize the cosine distance between embeddings of positive pairs (a, p), where both x_a and x_p are instances of eating gestures (e.g., using different utensils or eating at varying speeds). The corresponding loss is:

$$L_{\text{pos}} = D(f(x_a), f(x_p))^2, \tag{3}$$

which encourages feature consistency within the eating class. For negative pairs (a, n), where x_n is a non-eating activities (e.g., smoking or adjusting glasses), we apply:

$$L_{\text{neg}} = \max(\text{margin} - D(f(x_a), f(x_n)), 0)^2, \tag{4}$$

to ensure sufficient separation between eating and non-eating features. We use a margin of 1.0, a common choice in contrastive learning [52]. The final contrastive objective is the sum of L_{pos} and L_{neg}, promoting both intra-class compactness and inter-class separation.

VII. Multimodal Feature Fusion

Multimodal data integration, combining inertial and acoustic signals, enhances fine-grained food category classification and chewing/biting detection in real-world scenarios. Inertial signals capture hand motion dynamics, while acoustic signals encode auditory patterns related to food texture. By leveraging these complementary strengths, *DietWatch* address the limitations of unimodal systems, such as difficulties in distinguishing foods with overlapping features [13], [14], [28], [30], [31]. For instance, slurping sounds are common across various noodle dishes, but inertial data can differentiate them by capturing utensil usage and associated hand gestures.

Attention-based Fusion Framework. To achieve effective feature fusion, we propose an attention-based framework that integrates acoustic and inertial features, as shown in Figure 7. Initially, denoised acoustic signals are converted into time–frequency spectrograms using the Short-Time Fourier Transform (STFT) [53]. We use a window size of 2048 samples and a hop length of 512 samples to ensure sufficient frequency resolution for capturing discriminative spectral features related to biting and chewing dynamics. Temporal dependencies are captured through GRU-based feature extractors tailored to each modality. The resulting feature embeddings from the acoustic and inertial branches are then concatenated and fed into a joint attention mechanism that assigns relevance weights to informative temporal segments across both modalities. This enables the model to dynamically emphasize acoustic features when food textures vary significantly, or inertial features when hand gestures provide more discriminative cues for classification.

Loss Function Design and Optimization. To support effective attention-based multimodal fusion described above, we design a composite loss function that jointly optimizes three objectives: classification accuracy, attention diversity, and parameter regularization. In the following loss functions, N denotes the number of samples in the training batch, which ensures balanced gradient contributions.

The classification loss L_c ensures that the fused features remain discriminative for both food classification and chewing/biting event recognition:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i), \tag{5}$$

where y_i is the ground truth label and $\hat{y}_i$ is the predicted probability. To encourage diverse attention patterns, we introduce an attention regularization loss that promotes orthogonality in

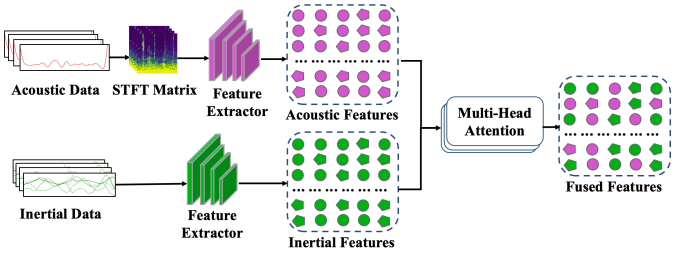


Fig. 7. Architecture of the attention-based multimodal feature fusion framework.

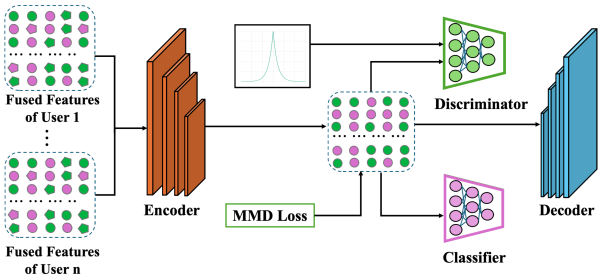


Fig. 8. Architecture of the cross-user adaptation framework.

the attention weight matrix and prevents collapse into a single focus:

$$L_{\text{att}} = \frac{1}{N}\sum_{i=1}^{N} \|AA^T - I\|_F, \tag{6}$$

where A is the attention weight matrix, I is the identity matrix, and $\|\cdot\|_F$ denotes the Frobenius norm. Additionally, a regularization term penalizes large model parameters to prevent overfitting:

$$L_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N} \|\theta\|_2, \tag{7}$$

where θ includes all trainable parameters of the model. The total loss function combines these objectives:

$$L_{\text{fusion}} = L_c + \lambda_p L_{\text{att}} + \lambda_q L_{\text{reg}}, \tag{8}$$

where λ_p and λ_q are empirically chosen weights to balance the three terms. After training, the learned feature extractors and attention modules are retained to enable robust fusion of real-time inertial and acoustic inputs.

VIII. Cross-user Adaptation

As discussed in Section III-B, user-specific variations in dietary behaviors, such as differences in eating sounds and hand gestures, pose significant challenges for model generalization. These variations often cause domain shifts, leading to performance degradation when applying pre-trained models to unseen users in real-world scenarios [54].

Adversarial Autoencoder Framework. To enable cross-user adaptation, we propose an adversarial autoencoder framework to learn user-invariant feature representations of eating behaviors, as illustrated in Figure 8. Our framework comprises four key components: (1) an encoder network that transforms the fused multi-modal features into a latent representation; (2) a decoder network that reconstructs the original fused features to preserve essential behavioral information; (3) a discriminator that enforces a Laplace prior over the latent space to promote sparsity, encouraging the model to isolate salient behavioral patterns and improve generalization; and (4) a domain adaptation module based on Maximum Mean Discrepancy (MMD), which explicitly aligns the latent feature distributions across users.

Loss Function Design and Optimization. The training process jointly optimizes three objectives: classification accuracy, feature reconstruction, and representation regularization. The classification loss L_{c} ensures that the fused features are discriminative for food classification and chewing/biting detection. The feature reconstruction loss is defined as:

$$L_r = \frac{1}{N}\sum_{i=1}^{N} \text{MSE}(p_i, \hat{p}_i), \tag{9}$$

where N is the batch size used during training, p_i and $\hat{p}_i$ denote the original and reconstructed feature vectors, and MSE represents the Mean Squared Error. To regularize the latent space, we adopt a Laplace prior, which promotes sparsity by assigning higher probability to values near zero. This helps the model isolate distinctive features relevant to dietary behaviors. The adversarial loss is formulated as:

$$L_a = \frac{1}{N}\sum_{i=1}^{N} \text{MSE}(h_i, l_i), \tag{10}$$

where h_i is the predicted latent representation and l_i is a sample from the Laplace prior. The MMD loss explicitly aligns latent feature distributions across users to mitigate user shifts:

$$L_m = \max\left(\sum_{u,v}\left\|\frac{1}{N_u}\sum_{i=1}^{N_u} E(p_{u,i}) - \frac{1}{N_v}\sum_{i=1}^{N_v} E(p_{v,i})\right\|, 0\right), \tag{11}$$

where $E(p_{u,i})$ and $E(p_{v,i})$ represent the latent representation for different users. The total loss functions are defined as:

$$L_{\text{adapt}} = L_{\text{c}} + \lambda_a L_{\text{r}} + \lambda_b L_{\text{a}} + \lambda_c L_{\text{m}}, \tag{12}$$

where λ_a, λ_b, and λ_c are empirically tuned weighting parameters to balance optimization and improve convergence. During training, *DietWatch* leverages fused features from the multi-modal feature fusion module to train the encoder, decoder, and discriminator collaboratively. In the inference phase, only the trained encoder is retained to extract user-invariant latent representations. This enables robust dietary monitoring for previously unseen users without additional training efforts.

IX. Fine-grained Dietary Behavior Derivation

A. Eating Time Derivation

To derive eating time intervals, we analyze the temporal distribution of identified eating gestures produced by the gesture recognition module. A temporal clustering algorithm is applied to group temporally adjacent gestures into coherent eating periods. The inertial data stream is segmented into 3-second windows, each independently classified as either eating or non-eating. Each recognized gesture g_i is associated with a timestamp t_i, forming a time-ordered sequence $G = g_1, g_2, \ldots, g_n$. An eating period C_k is defined as a consecutive subsequence of gestures satisfying:

$$C_k = \{g_i, g_{i+1}, \ldots, g_j \mid t_{m+1} - t_m < \theta,\ \forall m \in [i, j-1]\}. \tag{13}$$

Fig. 9. Real-life eating scenarios for data collection.

Based on typical meal patterns and short pauses during eating, we empirically set θ to 25 seconds [22]. Gestures separated by larger gaps are assigned to different periods. To improve robustness, we retain only eating periods that contain at least 4 gestures per minute and last no less than 3 minutes.

B. Food Type Classification

Both food type classification and biting/chewing frequency estimation are performed within the identified eating periods. The classifier takes as input the latent representations produced by the cross-user adaptation module and outputs food category predictions. For food type classification, we employ a neural network comprising two fully connected layers with ReLU activation, followed by a softmax output layer that generates a probability distribution over C predefined food categories. The model is trained using a standard cross-entropy loss function to maximize classification accuracy. During inference, we adopt a confidence-based decision rule: a prediction is accepted only if the highest class probability exceeds a threshold of 0.85. This strategy helps suppress low-confidence predictions arising from noisy or ambiguous inputs.

C. Biting/Chewing Frequency Estimation

To achieve a comprehensive analysis of eating behaviors, our system also estimates both biting and chewing frequency over the eating period. We employ deep learning-based classifiers to detect biting and chewing events within each sliding windows (e.g., 3 seconds). Specifically, we apply a binary classifier $\mathbb{I}_b(\cdot)$ to identify whether a bite occurs within each window, and a multi-class classifier $\mathbb{I}_c(\cdot)$ to predict the number of chewing actions (ranging from 0 to 6) within each window. These per-window predictions are then aggregated over the eating period to compute the overall biting and chewing frequencies.

$$f_{\text{bite}} = \frac{\sum_j \mathbb{I}_b(|v_{\text{latent}}(j)|)}{T_i}, f_{\text{chew}} = \frac{\sum_j \mathbb{I}_c(|v_{\text{latent}}(j)|)}{T_i}, \quad (14)$$

Here, $v_{\text{latent}}(j)$ denotes the latent features at the jth sliding window, and T_i is the duration of the eating period.

X. Performance

Device: We develop a prototype of DietWatch on three mainstream smartwatches: the Samsung Galaxy Watch 5, Google Pixel Watch 2, and Apple Watch Series 9. Each device integrates a 6-axis IMU (3-axis accelerometer and 3-axis gyroscope) and a microphone sampled at 41 kHz.

Experimental Scenarios: The system is evaluated across four representative real-life dining contexts with distinct types of dynamic interference, as illustrated in Figure 9(a):

TABLE II
Activity Categories in Dataset

Category	Activity Types
Eating Activities: 7 types	Using Spoon, Using Fork, Using Chopsticks, Drinking from Cup, Drinking from Bottle, Cutting Food, Eating by Hand
Non-eating Activities: 16 types	Phone Call, Standing, Shaving Face, Cleaning Ears, Conversation, Whistling, Nail Biting, Teeth Brushing, Sitting, Head Scratching, Teeth Picking, Hair Combing, Walking, Keyboard Typing, Book Reading, Phone Browsing

1) Low Disturbance Dining (LDD): Participants eat alone in a quiet room (less than 30 dB), without walking or multitasking. *2) Public Venue Dining (PVD):* Participants eat in a coffee shop with moderate background noise from machines (65–70dB), music (50–55dB), and conversations (55–60dB). *3) Mobile Context Dining (MCD):* Participants eat portable food (e.g., cookies, fruits) while walking indoors at 4–5 km/h, introducing continuous motion and ambient noise. *4) Social Dining (SD):* Participants eat with 1–3 others while engaging in natural conversations (60–65 dB) and frequent hand gestures (e.g., phone use).

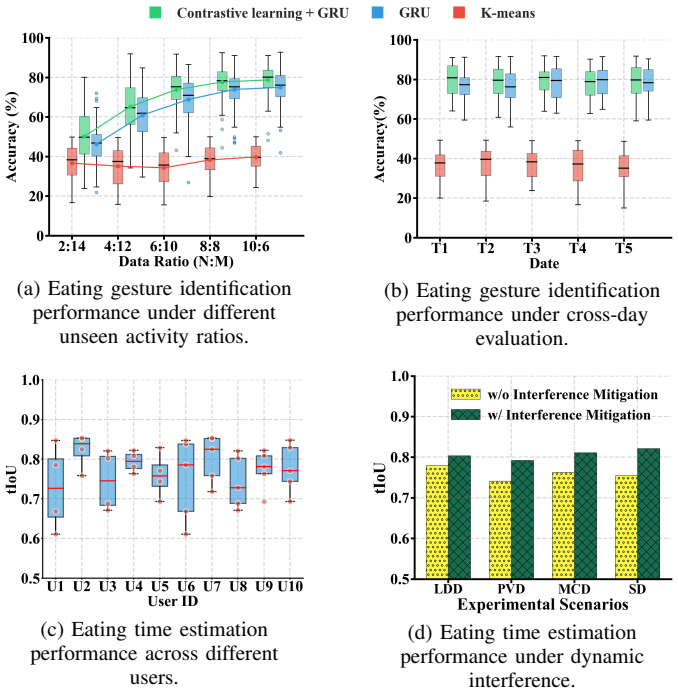
Data Collection: We conducted experiments with 30 participants (17 males, 13 females), aged 20–59, each wearing a DietWatch-enabled smartwatch on their dominant wrist. Participants performed eating and non-eating activities under the four defined real-life dining scenarios. All participants provided informed consent, and the experimental protocol was approved by the Institutional Review Board of Yeshiva University. The data were collected over a period of four months, resulting in more than 500 eating sessions and over 5500 minutes of recordings. Each session lasted 10 minutes and included both eating and non-eating activities. The experiments include 7 types of eating activities and 16 types of non-eating activities as shown in Table II. A total of 40 food types spanning 5 categories—staple foods, hard/crispy foods, soft foods, fruits/vegetables, and beverages are evaluated in the experiments and illustrated in the Table III.

Ground-truth annotations were obtained via synchronized video recordings using a GoPro Hero 12 camera. To ensure annotation reliability, we employed a semi-automated validation process: the Python script first provided automated timestamp suggestions based on video analysis, which were then manually verified and refined by the researcher through careful frame-by-frame inspection. Ambiguous cases, such as unclear gesture boundaries or overlapping activities, were systematically reviewed through repeated playback.

Evaluation Metrics: *1) Accuracy.* Accuracy is used to evaluate the performance of the eating gesture identification and food type classification modules. It represents the proportion of samples correctly predicted as the true labels (e.g., eating gesture or food type). *2) Temporal Intersection over Union (tIoU).* For eating time estimation, we adopt temporal Intersection over Union (tIoU) to measure the alignment between predicted and ground-truth eating periods. It is defined as: $tIoU = \frac{|D \cap G|}{|D \cup G|}$, where D and G denote the predicted and ground-truth eating periods, respectively. A higher tIoU indicates better tem-

TABLE III
FOOD CATEGORIES (EACH CATEGORY CONTAINS 8 FOOD TYPES)

Category	Food Types
(a) Staple Foods	Rice, Corn, Bread, Crackers, Fried buns, Boiled potatoes, Potato, Cereal
(b) Hard/Crispy Foods	Chips, Cookies, Fries, Peanut, Pecans, Gum, Chocolate, M&M's
(c) Soft Foods	Yogurt, Pudding, Cake, Egg, Ice cream, Mousse, Marshmallow, Meat
(d) Fruits/Vegetables	Apple, Pear, Orange, Tangerine, Grape, Carrot, Tomato, Cucumber
(e) Beverages	Water, Tea, Coffee, Milk, Juice, Cola, Wine, Parsley



(a) Eating gesture identification performance under different unseen activity ratios.

(b) Eating gesture identification performance under cross-day evaluation.

(c) Eating time estimation performance across different users.

(d) Eating time estimation performance under dynamic interference.

Fig. 10. Performance evaluation of eating gesture identification and eating time estimation.

poral alignment and more accurate eating time estimation. *3) Mean Absolute Error (MAE).* We assess the performance of chewing/biting frequency estimation using MAE, which measures the average absolute difference between the predicted and ground-truth frequencies (in times/min). It is calculated as $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_{\text{true},i} - f_{\text{estimated},i}|$, where n is the number of eating periods, $f_{\text{true},i}$ is the ground truth frequency for the i-th eating period, and $f_{\text{estimated},i}$ is our estimated frequency for the i-th eating period.

A. Performance of Eating Time Estimation

Robustness to Unseen Non-Eating Activities. To evaluate the robustness of our eating gesture recognition approach against unseen non-eating activities, we design experiments by varying the number of non-eating activity types in the training set from 2 to 10 (out of a total of 16). For each configuration, the remaining categories (from 14 to 6) are used exclusively in the testing set to simulate unseen scenarios. We conduct 50 repetitions per configuration using randomly sampled category subsets to ensure statistical reliability and minimize the impact of selection bias. We compare our method against two baselines: (1) the same model without incorporating the contrastive feature alignment, and (2) an unsupervised k-means clustering approach [55]. As shown in Figure 10(a), our method consistently outperforms both baselines across all configurations. Notably, even under the most challenging setting, where the training set includes only 2 non-eating activity categories and the testing set contains 14, our approach achieves an identification accuracy of 82.40%.

Robustness to Behavioral and User Variations. To evaluate the system's ability to extract consistent intake gesture features across different scenarios, emotional states, and individual habits, we conduct two sets of experiments: cross-day and cross-user evaluations. In the cross-day evaluation, data are collected from five separate days spaced two weeks apart, introducing natural variability in behavior such as differences in speed, motion range, and daily conditions. To evaluate robustness, we employ a cross-validation strategy that uses data from one day as the testing set and data from another day as the training set. As shown in Figure 10(b), the x-axis denotes the selected testing day (T1–T5). Our system achieves consistently high eating gesture identification accuracy across all test days, with median performance exceeding 80%, demonstrating strong resistance to day-to-day variation.

In the cross-user setting, we evaluate how well the system performs eating time estimation when applied to previously unseen individuals. We first establish a within-user baseline, where training and testing are both conducted on data from the same participant. In this setting, our system achieves an average tIoU of 77.79%. We then perform leave-one-out cross-validation (LOOCV) on data from 10 participants. In each iteration, the model is trained on data from 9 users and tested on the remaining one. This setup reflects practical deployment conditions where user-specific data may be unavailable. As shown in Figure 10(c), the x-axis represents the testing user ID (U1–U10). The results show that our system maintains comparable performance across all users, with median tIoU consistently above 76%, validating the robustness of the extracted features against inter-user differences.

Robustness to Dynamic Interference. To assess the system's robustness in realistic and dynamically changing dining environments, we evaluate the performance of eating time estimation across four representative scenarios. As shown in Figure 10(d), without the interference mitigation module, the system achieves a tIoU of 77.96% in low disturbance dining (LDD), 74.12% in public venue dining (PVD), 76.23% in mobile context dining (MCD), and 75.45% in social dining (SD). With the interference mitigation module enabled, performance improves consistently across all scenarios, with each bar in Figure 10(d) showing a marked increase. On average, the tIoU increases by 4.75% across all scenarios, reaching an overall accuracy of 79.95%. This validates the effectiveness of our interference mitigation module in preserving accurate eating period boundaries, even in complex real-world scenarios.

B. Performance of Food Type Classification

Overall Performance. To evaluate our system's ability to classify food types, we selected 40 commonly consumed food

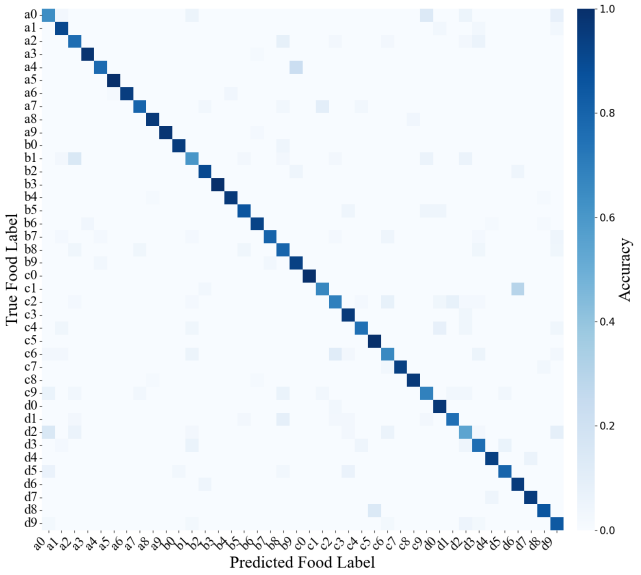
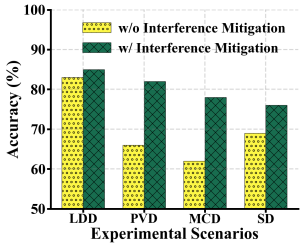


Fig. 11. Confusion matrix of food type classification (40 types across 5 categories; a–e correspond to Table III).

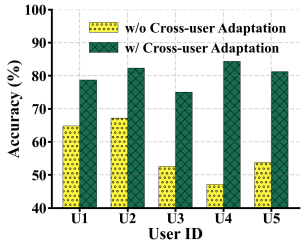
types as listed in Table III. We compare the proposed multimodal feature fusion framework with two baseline models: one using only inertial features and the other using only acoustic features. The inertial baseline achieves an average accuracy of 35.6%, while the acoustic baseline achieves 52.4%. In contrast, our proposed method achieves a significantly higher accuracy of 85.68%, demonstrating the effectiveness of feature fusion.

Figure 11 shows the confusion matrix across all 40 food types. The system performs particularly well on hard/crispy foods, achieving 90.85% accuracy, followed by staple foods (86.25%), fruits/vegetables (85.12%), soft foods (84.35%), and beverages (83.25%). The system maintains clear category boundaries, with a low cross-category confusion rate of 0.42% (i.e., the percentage of misclassifications where the predicted food belongs to a different major category). In contrast, the average intra-category confusion rate is 14.38% (i.e., misclassifications within the same major category), indicating that most classification errors occur between similar food types. blueDetailed examination of the confusion matrix reveals that these intra-category misclassifications occur primarily between foods that exhibit both similar acoustic signatures and comparable eating gestures. For instance, soft foods like yogurt and pudding share similar acoustic properties due to their viscosity. However, our multimodal approach can often distinguish them through differences in eating gestures (e.g., scooping vs. drinking). When both acoustic and inertial features are similar—such as between crispy foods like chips and cookies, citrus fruits like oranges and tangerines, or nuts like peanuts and pecans—misclassifications still occur. These results demonstrate that our multimodal fusion effectively leverages complementary cues from both modalities, though foods sharing similar characteristics across both domains remain challenging to distinguish. These results suggest that our system effectively distinguishes among high-level food categories while maintaining strong type-level resolution.

Robustness to Dynamic Interference. To evaluate the



(a) Food type classification performance under dynamic interference.



(b) Food type classification performance across different users.

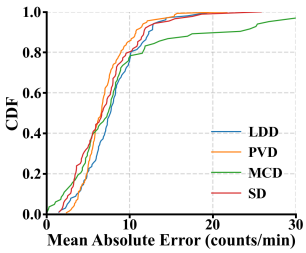
Fig. 12. Performance of food type classification in real-world scenarios.

robustness of food type classification in real-world conditions, we assess the system's performance across four representative scenarios. As illustrated in Figure 12(a), without interference mitigation module, the system achieves accuracies of 83% in low-disturbance dining (LDD), 66% in public venues (PVD), 62% in mobile contexts (MCD), and 69% in social dining (SD). With the interference mitigation module enabled, the system consistently improves across all scenarios, with accuracy gains exceeding 10% in most cases. These results highlight the module's effectiveness in preserving essential eating features and alleviating the impact of dynamic disturbances in real-world settings.

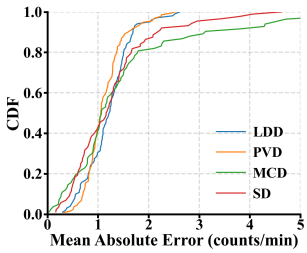
Robustness to User Variations. To assess the system's generalization ability across users in food type classification, we conduct a cross-user experiment involving five participants. In each test, the model is trained on data from four users and evaluated on the fifth, previously unseen, user. For baseline comparison, we also perform within-user testing, where both training and testing data come from the same user. Without applying our cross-user adaptation framework, the system reveals a notable performance drop, with an average accuracy decreasing by 29.8% compared to within-user testing. This result highlights the challenge posed by individual variability in food type classification. As illustrated in Figure 12(b), our cross-user adaptation framework significantly mitigates this drop. After incorporating our adaptation framework, the classification accuracy improves markedly for all users, with an average gain of over 23.2%. These results demonstrate the system's ability to adapt effectively to new users and mitigate the performance degradation caused by user diversity.

C. Performance of Biting and Chewing Frequency Estimation

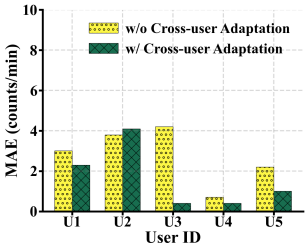
Robustness to Dynamic Interference. We evaluate our system's performance in estimating chewing and biting frequencies under 4 real-world scenarios. Participants are instructed to chew and bite food at normal rates, approximately 45 chews per minute and 5 bites per minute, respectively [20], [56]. Video recordings are used to establish the ground truth for evaluation. As shown in Figure 13(a), the cumulative distribution function (CDF) of the mean absolute error (MAE) in chewing frequency estimation demonstrates robust performance across all scenarios. In LDD scenario, over 80% of the MAE values fall below 7.3 chews/min. The system maintains consistent accuracy in PVD, MCD, and SD scenarios, with over 80% of the MAE values remaining below 6.7 chews/min



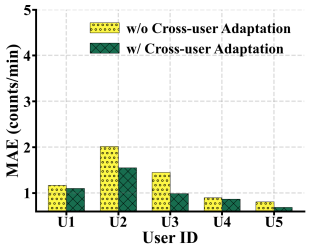
(a) Performance of chewing frequency estimation under dynamic interference.



(b) Performance of biting frequency estimation under dynamic interference.

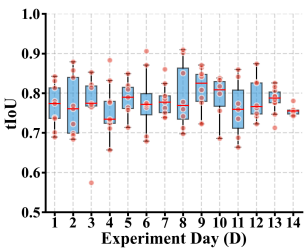


(c) Performance of chewing frequency estimation across different users.

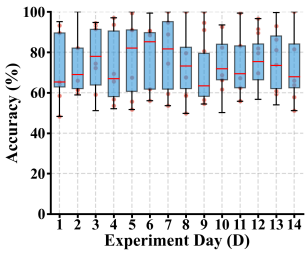


(d) Performance of biting frequency estimation across different users.

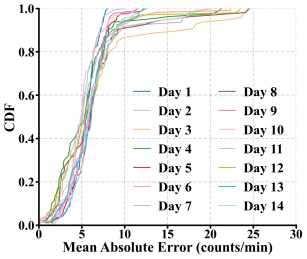
Fig. 13. (a–b) Cumulative distribution functions (CDF) of the mean absolute error (MAE) for chewing and biting frequency estimation under 4 real-world scenarios. (c–d) MAE comparison across different users before and after applying the cross-user adaptation module for chewing and biting frequency estimation.



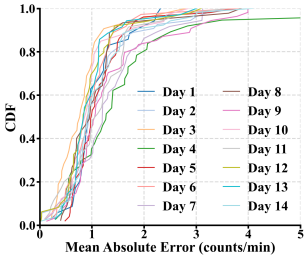
(d) Eating time estimation.



(a) Food type classification.



(b) Chewing frequency estimation.



(c) Biting frequency estimation.

Fig. 14. Performance of DietWatch in real-world environments over a 14-day study.

in each case. Figure 13(b) shows the corresponding results for biting frequency estimation. In LDD, over 80% of the MAE is below 1.5 bites/min. In PVD, the same proportion of errors falls below 1.3 bites/min. The system also sustains comparable accuracy in MCD and SD, with over 80% of the MAE below 1.6 bites/min across both. These results demonstrate that our system achieves accurate chewing and biting frequency estimation across diverse real-world scenarios.

Robustness to User Variations. We collected data from five participants to evaluate the robustness of our frequency estimation approach under user variations. In the within-user testing setting, our system achieves an average MAE of 7.71 counts/min for chewing frequency estimation and 1.26 counts/min for biting frequency estimation. To assess generalization to unseen users, we conduct LOOCV. Without cross-user adaptation framework, the estimation performance degrades significantly compared to the within-user setting: the MAE increases by 17% for chewing frequency and 21% for biting frequency. With the proposed cross-user adaptation, the performance improves markedly. As shown in Figures 13(c) and (d), the adaptation module effectively reduces the MAE across users. Specifically, the median MAE for chewing frequency decreases from 8.99 to 6.99 counts/min, while that for biting frequency decreases from 1.52 to 1.11 counts/min.

XI. Real World Case Study

To further evaluate *DietWatch*'s practical application capability in real-world scenarios, we conduct a 14-day real-world tracking study. This study involves 5 participants who use *DietWatch* during their daily lives over 14 consecutive days. Participants are required to wore smartwatches for continuous data collection and record videos as ground truth references. The study collects approximately 70 hours of data across diverse real-world environments, including homes, offices, restaurants, cafes, and bedrooms. To evaluate robustness beyond controlled settings, participants engaged in various complex dining scenarios such as communal dining with family members and social gatherings with 3 people. Participants consumed culturally diverse cuisines such as curry, dumplings, kung pao chicken, pasta, and traditional American dishes. Figure 14 summarizes the system's performance across multiple tasks over the 14-day period.

Figure 14(a) shows the performance of eating moment estimation across the 14-day study. Our system consistently achieved high tIoU scores, with daily medians ranging from approximately 0.69 to 0.80. The interquartile ranges remain narrow on most days, indicating stable performance with limited variance. Figure 14(b) presents the food type classification performance. Our system maintains stable performance across different days, with average accuracy around 73%. While some fluctuations are observed such as Day 4 and Day 7, the overall results demonstrate that our system remains robust under real-world scenarios. Figure 14(c) shows the results of chewing frequency estimation. Overall, the system achieves consistent performance, with 85% estimations yielding MAE within 10 counts/min. Figure 14(d) presents the system's performance in estimating biting frequency. The system also maintains stable performance throughout the period, with 68% to 85% of the estimations each day achieving a MAE below 1.5 counts/min. Overall, the consistent performance across all days validates the effectiveness of our approach in real-world scenarios.

XII. Discussion

Applications Enabled by DietWatch. *DietWatch* leverages commercial smartwatches to collect data, offering a widely accessible and scalable approach to unobtrusive dietary monitoring. It provides retrospective insights into users' eating

behaviors, allowing them to reflect on dietary habits and trends. By supporting longitudinal tracking of eating patterns, *DietWatch* is well-suited for applications such as habit formation analysis and personalized behavioral coaching. Looking ahead, we plan to enhance *DietWatch*'s capabilities to estimate caloric intake and nutritional composition by integrating fine-grained food composition databases. This advancement could enable more tailored dietary guidance, including daily caloric targets and macronutrient balance optimization.

Limitations and Future Work. *DietWatch* currently relies on continuous IMU sampling to capture motion data. To conserve battery and protect user privacy, the microphone is triggered only when an eating event is detected. This design helps reduce power consumption and mitigates potential privacy concerns associated with constant audio recording.

The system currently performs deep learning inference on a remote server. While effective, server-side inference introduces network dependencies and may pose challenges in privacy-sensitive environments. To reduce reliance on cloud resources, we plan to migrate early-stage processing modules—such as interference mitigation and feature extraction—to the smartwatch.

To support this migration, we will apply model compression techniques such as quantization and pruning [57] to develop lightweight version of interference mitigation modules that can operate efficiently on wearable hardware. We will further explore split learning [58] to allow the early portions of the feature extractors to run locally on the smartwatch while transmitting only intermediate representations to the server. Additionally, federated learning [59] offers a promising avenue for *DietWatch* to enable more personalized dietary monitoring while ensuring that raw data remains local to individual devices. Collectively, these approaches provide the technical foundation needed to move portions of the *DietWatch* pipeline onto the device and decrease dependence on remote inference.

Beyond deployment considerations, there is also room to improve classification accuracy through more advanced multimodal fusion architectures. Transformer-based fusion models [60] and lightweight temporal attention mechanisms [61] may help reduce intra-class confusion and better leverage cross-modal dependencies. We plan to explore these algorithmic refinements alongside enhanced privacy-aware deployment strategies in future system iterations.

Finally, while *DietWatch* performs reliably in typical social dining environments where individuals are seated at least 1.2 meters apart, challenges remain in close-proximity scenarios. When two users sit shoulder-to-shoulder and share snacks, overlapping acoustic signals—arising from simultaneous eating sounds and nearby conversations—introduce complex interference that exceeds the capability of our current denoising approach. Addressing this limitation will require future development of advanced source separation algorithms to isolate eating-related sounds in such environments.

XIII. Conclusion

In this paper, we propose *DietWatch*, a fine-grained dietary monitoring system designed for real-world scenarios using a commercial smartwatch. By addressing challenges such as dynamic interference, gesture generalization, and user diversity, *DietWatch* can provide robust dietary behavior analysis with minimal user effort. Experimental results demonstrate the system's effectiveness, achieving 79.75% tIoU for eating time estimation, 86.0% accuracy in food type classification, and MAE of 1.2 bites/min for biting frequency and 7.8 chews/min for chewing frequency estimation. *DietWatch* offers a scalable and practical solution for dietary monitoring, paving the way for advancements in personalized nutrition and public health interventions.

References

[1] World Health Organization, "Healthy diet," 2020, accessed: 2023-09-30. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/healthy-diet

[2] R. M. Leech, A. Timperio, K. M. Livingstone, A. Worsley, and S. A. McNaughton, "Temporal eating patterns: associations with nutrient intakes, diet quality, and measures of adiposity," *The American journal of clinical nutrition*, vol. 106, no. 4, pp. 1121–1130, 2017.

[3] Y. Zhu and J. H. Hollis, "Increasing the number of chews before swallowing reduces meal size in normal-weight, overweight, and obese adults," *Journal of the Academy of Nutrition and Dietetics*, vol. 114, no. 6, pp. 926–931, 2014.

[4] M. J. Gibney and M. C. Walsh, "The future direction of personalised nutrition: my diet, my phenotype, my genes," *Proceedings of the Nutrition Society*, vol. 72, no. 2, pp. 219–225, 2013.

[5] J.-S. Shim, K. Oh, and H. C. Kim, "Dietary assessment methods in epidemiologic studies," *Epidemiology and health*, vol. 36, 2014.

[6] F. E. Thompson and A. F. Subar, "Dietary assessment methodology," *Nutrition in the Prevention and Treatment of Disease*, pp. 5–48, 2017.

[7] S. Bi, T. Wang, N. Tobias, J. Nordrum, S. Wang, G. Halvorsen, S. Sen, R. Peterson, K. Odame, K. Caine *et al.*, "Auracle: Detecting eating episodes with an ear-mounted sensor," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–27, 2018.

[8] V. Papapanagiotou, C. Diou, J. van den Boer, M. Mars, and A. Delopoulos, "Recognition of food-texture attributes using an in-ear microphone," in *International Conference on Pattern Recognition*. Springer, 2021. [Online]. Available: https://link.springer.com/conference/icpr/your-specific-document-id

[9] J. Qiu, F. P.-W. Lo, and B. Lo, "Assessing individual dietary intake in food sharing scenarios with a 360 camera and deep learning," in *2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2019, pp. 1–4.

[10] M. Sun, L. E. Burke, Z.-H. Mao, Y. Chen, H.-C. Chen, Y. Bai, Y. Li, C. Li, and W. Jia, "ebutton: a wearable computer for health monitoring and personal assistance," in *Proceedings of the 51st annual design automation conference*, 2014, pp. 1–6.

[11] J. Liu, E. Johns, L. Atallah, C. Pettitt, B. Lo, G. Frost, and G.-Z. Yang, "An intelligent food-intake monitoring system using wearable sensors," in *2012 ninth international conference on wearable and implantable body sensor networks*. IEEE, 2012, pp. 154–160.

[12] S. A. Rahman, C. Merck, Y. Huang, and S. Kleinberg, "Unintrusive eating recognition using google glass," in *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE, 2015, pp. 108–111.

[13] S. Bi and D. Kotz, "Eating detection with a head-mounted video camera," in *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2022, pp. 60–66.

[14] M. Mirtchouk, D. Lustig, A. Smith, I. Ching, M. Zheng, and S. Kleinberg, "Recognizing eating from body-worn sensors: Combining free-living and laboratory data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–20, 2017.

[15] S. Sharma, P. Jasper, E. Muth, and A. Hoover, "The impact of walking and resting on wrist motion for automated detection of meals," *ACM Transactions on Computing for Healthcare*, vol. 1, no. 4, pp. 1–19, 2020.

[16] M. Farooq, J. M. Fontana, and E. Sazonov, "A novel approach for food intake detection using electroglottography," *Physiological measurement*, vol. 35, no. 5, p. 739, 2014.

[17] P. Tseng, B. Napier, L. Garbarini, D. L. Kaplan, and F. G. Omenetto, "Functional, rf-trilayer sensors for tooth-mounted, wireless monitoring of the oral cavity and food consumption," *Advanced Materials*, vol. 30, no. 18, p. 1703257, 2018.

[18] Demand Sage, "Smartwatch statistics 2024: Worldwide market data," 2024, [Online; accessed Day-Month-Year]. [Online]. Available: https://www.demandsage.com/smartwatch-statistics

[19] ElectroIQ, "Smartwatch statistics by brands, revenue, and users," 2024, [Online; accessed Day-Month-Year]. [Online]. Available: https://electroiq.com/smartwatch-statistics-by-brands

[20] C. Wang, T. Kumar, W. De Raedt *et al.*, "Eating speed measurement using wrist-worn imu sensors towards free-living environments," *IEEE Journal of Biomedical and Health Informatics*, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/your-specific-document-id

[21] B. Wei, S. Zhang, X. Diao, Q. Xu, Y. Gao, and N. Alshurafa, "An end-to-end energy-efficient approach for intake detection with low inference time using wrist-worn sensor," *IEEE Journal of Biomedical and Health Informatics*, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/your-specific-document-id

[22] S. Stankoski, M. Jordan, H. Gjoreski, and M. Luštrek, "Smartwatch-based eating detection: Data selection for machine learning from imbalanced data with imperfect labels," *Sensors*, vol. 21, no. 5, p. 1902, 2021.

[23] Y. Luktuke and A. Hoover, "Segmentation and recognition of eating gestures from wrist motion using deep learning," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/your-specific-document-id

[24] M. B. Morshed, S. S. Kulkarni, R. Li, K. Saha, L. G. Roper, L. Nachman, H. Lu, L. Mirabella, S. Srivastava, M. De Choudhury *et al.*, "A real-time eating detection system for capturing eating moments and triggering ecological momentary assessments to obtain further context: System development and validation study," *JMIR mHealth and uHealth*, vol. 8, no. 12, p. e20625, 2020.

[25] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[26] Z. Hou, Y. Xie, and F. Li, "Dietwatch: Towards low-effort fine-grained dietary monitoring via smartwatch in open-world scenarios," in *ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, ser. CHASE '25. New York, NY, USA: ACM, 2025, pp. 1–6.

[27] K. Yatani and K. N. Truong, "Bodyscope: a wearable acoustic sensor for activity recognition," in *Proceedings of the 2012 ACM conference on ubiquitous computing*, 2012, pp. 341–350.

[28] A. Bedri, D. Li, R. Khurana, K. Bhuwalka, and M. Goel, "Fitbyte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[29] A. Bedri, R. Li, M. Haynes, R. P. Kosaraju, I. Grover, T. Prioleau, M. Y. Beh, M. Goel, T. Starner, and G. Abowd, "Earbit: using wearable sensors to detect eating episodes in unconstrained environments," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 1, no. 3, pp. 1–20, 2017.

[30] S. Zhang, Y. Zhao, D. T. Nguyen, R. Xu, S. Sen, J. Hester, and N. Alshurafa, "Necksense: A multi-sensor necklace for detecting eating activities in free-living conditions," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 4, no. 2, pp. 1–26, 2020.

[31] R. Zhang, J. Zhang, N. Gade, P. Cao, S. Kim, J. Yan, and C. Zhang, "Eatingtrak: Detecting fine-grained eating moments in the wild using a wrist-mounted imu," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. MHCI, pp. 1–22, 2022.

[32] R. Zhang and O. Amft, "Monitoring chewing and eating in free-living using smart eyeglasses," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 23–32, 2017.

[33] F. Tehrani, H. Teymourian, B. Wuerstle, J. Kavner, R. Patel, A. Furmidge, R. Aghavali, H. Hosseini-Toudeshki, C. Brown, F. Zhang *et al.*, "An integrated wearable microneedle array for the continuous monitoring of multiple biomarkers in interstitial fluid," *Nature Biomedical Engineering*, vol. 6, no. 11, pp. 1214–1224, 2022.

[34] K. Kyritsis, C. Diou, and A. Delopoulos, "A data driven end-to-end approach for in-the-wild monitoring of eating behavior using smartwatches," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 1, pp. 22–34, 2020.

[35] E. Thomaz, I. Essa, and G. D. Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 2015, pp. 1029–1040.

[36] S. Sen, V. Subbaraju, A. Misra, R. Balan, and Y. Lee, "Annapurna: An automated smartwatch-based eating detection and food journaling system," *Pervasive and Mobile Computing*, vol. 68, p. 101259, 2020.

[37] Q. Huang, W. Wang, and Q. Zhang, "Your glasses know your diet: Dietary monitoring using electromyography sensors," *IEEE Internet of Things Journal*, vol. 4, no. 3, pp. 705–712, 2017.

[38] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Internet of things: Device capabilities, architectures, protocols, and smart applications in healthcare domain," *IEEE Internet of Things Journal*, vol. 10, no. 4, pp. 3611–3641, 2022.

[39] N. Mohammadian, A. Didikoglu, C. Beach, P. Wright, J. W. Mouland, F. P. Martial, S. Johnson, M. Van Tongeren, T. M. Brown, R. J. Lucas *et al.*, "A wrist-worn internet of things sensor node for wearable equivalent daylight illuminance monitoring," *IEEE Internet of Things Journal*, vol. 11, no. 9, pp. 16 148–16 157, 2024.

[40] H. Kalantarian and M. Sarrafzadeh, "Audio-based detection and evaluation of eating behavior using the smartwatch platform," *Computers in biology and medicine*, vol. 65, pp. 1–9, 2015.

[41] R. Zhang and O. Amft, "Retrieval and timing performance of chewing-based eating event detection in wearable sensors," *Sensors*, vol. 20, no. 2, p. 557, 2020.

[42] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[43] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover, "Detecting periods of eating during free-living by tracking wrist motion," *IEEE journal of biomedical and health informatics*, vol. 18, no. 4, pp. 1253–1260, 2013.

[44] S. Zhang, R. Alharbi, W. Stogin, M. Pourhomayun, B. Spring, and N. Alshurafa, "Food watch: Detecting and characterizing eating episodes through feeding gestures," in *Proceedings of the 11th EAI International Conference on Body Area Networks*, 2016, pp. 91–96.

[45] K. Cho, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[46] M. Tarek and S. Zahran, "Enhanced imu sensors fusion based on adaptive least square windowing technique," in *2021 International Telecommunications Conference (ITC-Egypt)*. IEEE, 2021, pp. 1–4.

[47] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[48] J. Xie, J. G. Colonna, and J. Zhang, "Bioacoustic signal denoising: a review," *Artificial Intelligence Review*, vol. 54, pp. 3575–3597, 2021.

[49] P. Hewage, A. Behera, M. Trovati, E. Pereira, M. Ghahremani, F. Palmieri, and Y. Liu, "Temporal convolutional neural (tcn) network for an effective weather forecasting using time-series data from the local weather station," *Soft Computing*, vol. 24, pp. 16 453–16 482, 2020.

[50] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 5220–5224.

[51] J. L. Scisco, E. R. Muth, and A. W. Hoover, "Examining the utility of a bite-count–based measure of eating activity in free-living human beings," *Journal of the Academy of Nutrition and Dietetics*, vol. 114, no. 3, pp. 464–469, 2014.

[52] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[53] J. B. Allen and L. R. Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[54] Y. Chang, A. Mathur, A. Isopoussu, J. Song, and F. Kawsar, "A systematic study of unsupervised domain adaptation for robust human-activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–30, 2020.

[55] T. M. Kodinariya, P. R. Makwana *et al.*, "Review on determining number of cluster in k-means clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.

[56] V. Papapanagiotou, C. Diou, L. Zhou, J. Van Den Boer, M. Mars, and A. Delopoulos, "A novel chewing detection system based on ppg, audio, and accelerometry," *IEEE journal of biomedical and health informatics*, vol. 21, no. 3, pp. 607–618, 2016.

[57] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman cod-

ing," in *International Conference on Learning Representations (ICLR)*, 2016.
[58] Z. Li, C. Yan, X. Zhang, G. Gharibi, Z. Yin, X. Jiang, and B. A. Malin, "Split learning for distributed collaborative training of deep learning models in health informatics," in *AMIA Annual Symposium Proceedings*, vol. 2023, 2024, p. 1047.
[59] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2017, pp. 1273–1282.
[60] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7077–7087.
[61] M. A. Khatun, M. A. Yousuf, S. Ahmed, M. Z. Uddin, S. A. Alyami, S. Al-Ashhab, H. F. Akhdar, A. Khan, A. Azad, and M. A. Moni, "Deep cnn-lstm with self-attention model for human activity recognition using wearable sensor," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 10, pp. 1–16, 2022.

Zhen Hou is a Ph.D. student in Health Informatics at the Luddy School of Informatics, Computing, and Engineering, Indiana University Indianapolis, Indianapolis. He received his M.S. degree in Applied Data Management and Analytics from Purdue University, Indianapolis, and his B.S. degree in E-Communication Engineering from Henan University, China. His research interests include health monitoring applications, healthcare data standardization, and health information exchange systems, with particular focus on smartwatch-based dietary monitoring and FHIR interoperability. His work has appeared in conferences including MedInfo and IEEE/ACM CHASE.

Yucheng Xie is an Assistant Professor in the Department of Graduate Computer Science and Engineering at the Katz School of Science and Health, Yeshiva University, New York. He received his Ph.D. in Electrical and Computer Engineering from Purdue University, Indianapolis. His research interests include mobile sensing and computing, and security in machine learning and AI systems. His work has appeared in leading conferences and journals and has received multiple Best Paper and Runner-up Awards.

Feng Li is a Professor with the Department of Computer and Information Technology, Purdue University in Indianapolis, Indianapolis, IN, USA. He received the Ph.D. degree in computer science from Florida Atlantic University, Boca Raton, FL, USA, in 2009. He is actively involved in interdisciplinary collaborations and industry partnerships to advance cybersecurity education and research. He has led multiple funded research projects and has published extensively in top-tier journals and conferences. His research focuses on cybersecurity, trustworthy AI, federated learning, and robust data analytics, with applications in security, critical infrastructure, and intelligent systems.

Chengyi Liu is a Ph.D. student in computer science at the Katz School of Science and Health, Yeshiva University, New York. She received her M.S. degree in computer science from Northeastern University in 2022, and her B.S. degree in math from University of Illinois at Urbana Champaign in 2020. Her research interests include mobile computing and sensing, human activity recognition, machine learning/artificial intelligence.

Honggang Wang is the founding Chair and Professor of the Department of Graduate Computer Science and Engineering, Katz School of Science and Health, Yeshiva University in New York City. He is an alumnus of NAE Frontiers of Engineering program. He produced high-quality publications in prestigious journals and conferences in his research areas, winning several prestigious best paper awards. He is an IEEE distinguished lecturer and a Fellow of IEEE and AAIA. He has served as the Editor in Chief (EiC) for IEEE Internet of Things Journal during 2020–2022. He was the past Chair (2018–2020) of IEEE Multimedia Communications Technical Committee and the past IEEE eHealth Technical Committee Chair (2020–2021).