E-M^2: Efficient Multimodal Sensing via Adaptive Sensor-Computation Activation

Jinyi Cui, *Student Member, IEEE*, Tianyue Zheng✉, *Member, IEEE*

Abstract—Multimodal sensing systems have gained widespread adoption in IoT and edge intelligence, due to their ability to collect more comprehensive information of the target, thus improving sensing accuracy. However, this improved accuracy often comes at the cost of increased power consumption and computational overhead. Specifically, sensors require continuous power to collect data, and data processing algorithms not only demand intensive computational resources, but also consume significant energy to analyze the data, hindering the deployment of such systems on the edge. To address this issue, we propose E-M^2, a framework for efficient multimodal sensing by adaptive sensor-computation activation. First, E-M^2 selectively disables redundant modalities and corresponding data processing modules, effectively reducing unnecessary power consumption and computational overhead. Second, E-M^2 employs an exploration mechanism to reactivate disabled modalities, thus preventing "dead" modalities and enhancing overall system utilization. Finally, E-M^2 conditions the data processing algorithms on the on/standby states of the modalities, thus alleviating the negative impacts of sensor deactivation. Extensive evaluations demonstrate that E-M^2 reduces average power consumption by 40.75% and computational overhead by 48.84% across various sensing tasks, all while maintaining the sensing performance.

Index Terms—Multimodal sensing, sensor-computation activation, power consumption, computational overhead.

1 Introduction

Edge intelligence is a critical component of the rapidly growing Internet of Things (IoT), playing a vital role in processing data closer to the source. In edge intelligence systems, sensors are employed to capture data from targets in the environments, which are then processed by data processing modules that run neural networks to interpret the data and produce insights. Edge intelligence is increasingly adopting a multimodal approach, spurred by rapid advancements in sensing technology, reduced sensor manufacturing costs, and the imperative to improve sensing performance through complementary information from multiple modalities. Research consistently demonstrates that the use of complementary data from various sensing modalities, can significantly improve sensing performance [1], [2], [3], [4]. Consequently, leading consumer electronics products, such as smartphones [5], [6], fitness trackers [7], [8], and smart home systems [9], [10] are now equipped with multiple modalities to improve functionalities such as environment sensing [11], [12], health monitoring [13], [14], and human-computer interactions [15], [16], making these devices smarter and more responsive to user needs.

Despite their improved accuracy, multimodal systems incur significant power and computational overhead, which can be attributed to i) multimodal systems that incorporate multiple sensors, each requiring energy-intensive data acquisition and processing, and ii) the additional computational resources that must be allocated for processing data from multiple modalities. Consider a typical multimodal sensing system consisting of an RGB camera, a depth sensor, and a radar. Compared with a unimodal system, this multimodal setup experiences a 2.63 x increase in power consumption and an 8.24 x increase in computational overhead. Unfortunately, this high power and computational overhead hinder the practical application of multimodal sensing systems in edge computing environments. For example, it is revealed that a lithium-polymer battery of 5000 mAh/3.7 V, commonly found in edge devices, can be depleted within 1.8 hours by continuous monitoring using a multimodal system [17]. Furthermore, a neural network for multimodal sensing can easily reach a peak memory usage of 10 GB, exceeding the typical memory limit of 8 GB available in edge environments [18]. Therefore, it is imperative for academia and industry to optimize the power and computational overhead of multimodal sensing and computing systems [19], [20].

One effective strategy to reduce system overhead is to actively remove redundant or underutilized modalities. Redundancy occurs when modalities overlap in their functionalities. For example, both the RGB camera and the depth sensor capture texture and contour information [21], [22], and both the depth sensor and the radar provide radial distance information [23], [24]. In LoS scenarios with ample lighting, RGB cameras alone can suffice to capture most information, thereby allowing for the deactivation of other sensors. Underutilized modalities, on the other hand, capture little useful information under specific conditions. In low-light scenarios, the passive nature of RGB cameras renders them ineffective, providing minimal data. Similarly, in occluded environments, both RGB cameras and depth sensors struggle to sense subjects accurately, making radar with penetration capability the preferred modality [25]. Furthermore, distant targets outside the detection range of RGB and depth cameras might require the use of radar with a longer detection range. Although we have highlighted several typical scenarios, it is impractical to design a set of rules

- *J. Cui and T. Zheng are with the Department of Computer Science and Engineering, Southern University of Science and Technology, China. E-mail: cuijy2024@mail.sustech.edu.cn, zhengty@sustech.edu.cn*
- ✉ *Corresponding author: Tianyue Zheng.*

that account for all unique environment conditions [26], [27].

Due to the impracticality of rule-based methods, researchers are increasingly adopting data-driven approaches [28], [29], [30] to automatically control sensing modalities. While they improve efficiency by adaptively activating computing modules (e.g., deep learning networks), they fall short of achieving optimal power and computational efficiency. This limitation stems from their narrow focus on computational components while leaving sensor hardware continuously active, with associated computing modules constantly processing sensor outputs. The separation between hardware and software control layers creates a fundamental inefficiency, as the energy consumption of always-on sensors often dominates the system's power budget. Moreover, this bifurcated approach fails to leverage the intrinsic relationship between sensing and computation, where intelligent sensor activation could significantly reduce downstream computational requirements. As such, it is urgent for us to come up with an efficient multimodal sensing framework adaptively activating both sensors and computing modules.

While the high-level goal is intuitive, designing a practical framework for data-driven, joint hardware-software optimization poses significant challenges that are not found in single-sided optimization. First, the complex interdependency between sensing and computing requires a sophisticated decision-making mechanism. This mechanism should control both components simultaneously while adapting to changing environment conditions and system states. Second, deactivated sensing modalities create a critical reactivation problem: once a sensor enters standby, the system lacks the information necessary to determine when it should be reactivated, potentially leading to permanently inactive sensors and suboptimal utilization of available sensing capabilities. Third, the intermittent operation of sensing modalities introduces anomalous values into the computing modules, such as zeros or nulls that are not encountered during standard training, causing unpredictable outputs and compromising the reliability of the entire framework [31].

To address these challenges, we propose a framework for efficient multimodal sensing and computing, E-M^2, as shown in Fig. 1. E-M^2 dynamically deactivates non-essential sensing and computing modules at runtime, thus adapting to a variety of environments (e.g., daylight, low-light, occlusions, and distant targets) while achieving optimal energy and computational efficiency. E-M^2 employs a lightweight differentiable dual-pronged policy network (DPPN) for selecting sensing modalities and computing modules. DPPN is trained using normal backpropagation [32] alongside the sensing network itself, thereby avoiding a complicated rule-based design. To further enhance performance, E-M^2 employs long short-term temperature scheduling to explore the reactivation of deactivated modalities, maximizing overall sensing performance and ensuring rapid adaptation to ever-changing environments. Additionally, to address abnormal values (e.g., 0's and nulls) caused by deactivated modalities, E-M^2 employs a modality-conditioned training strategy. It explicitly informs the deep learning network about the availability of various modalities, ensuring the network is well-prepared to manage these anomalies during inference. In summary, our major contributions in this paper are as follows:

- To the best of our knowledge, E-M^2 is the first unified, data-driven framework for efficient multimodal sensing via adaptive sensor-computation activation.
- We design a unique DPPN to control the on/standby states of sensing and computing modules, adapting to ever-changing environments for optimal performance.
- We design a long short-term temperature scheduling to explore the reactivation of deactivated modalities, thereby maximizing overall modality utilization.
- We design a modality-conditioned training strategy to ensure that the computing modules are resilient to anomalies caused by deactivated modules.
- We implement E-M^2 prototype and evaluate E-M^2 with extensive experiments. The promising results demonstrate that E-M^2 can enable efficient multimodal sensing and computing.

The rest of the paper is organized as follows. § 2 motivates the design of E-M^2 by revealing the excessive power consumption and computational overhead associated with multimodal systems, as well as the redundancy inherent in their multiple modalities. § 3 presents the system design of E-M^2. § 4 introduces the datasets, system implementation, and experiment setup. § 5 reports the evaluation results in various scenarios. Finally, § 6 concludes the paper.

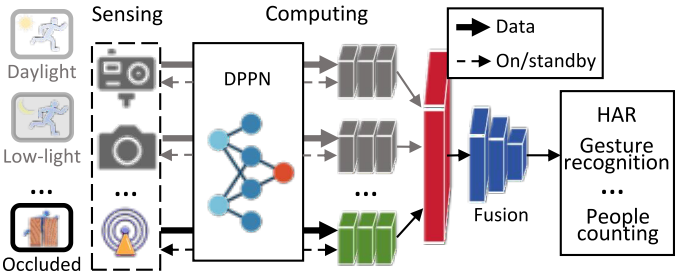


Fig. 1: E-M^2 employs a policy network to dynamically control the on/standby states of sensing and computing modules, adapting to ever-changing environments.

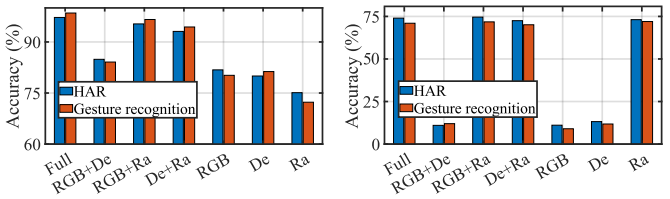
2 Motivation

In this section, we explain E-M^2's motivation. First, we demonstrate the existence of modality redundancy[1] and underutilization in multimodal systems. Next, we highlight the ever-changing nature of the sensing environment, underscoring the need for an adaptive framework to manage the on/standby states of the modalities and turn on deactivated modalities. Finally, we show that deactivating certain modalities can introduce abnormal values, which motivates the design of a modality-conditioned training strategy.

2.1 Modality Redundancy and Underutilization

A multimodal sensing system often experiences the issues of modality redundancy and underutilization. For our analysis, we consider a representative one with an RGB camera,

1. The concept of "modality redundancy" (the cause, where modalities provide overlapping information) should be distinguished from modality collapse [33], a potential result. The latter is an undesirable training-phase pathology where the model learns to ignore a modality.



(a) Daylight. (b) Occluded.

Fig. 2: Accuracy of different modality combinations in daylight and occluded environments, in which modality redundancy and underutilization occur

.

a depth sensor (hereafter, De), and a radar (hereafter, Ra). Fig. 2(a) illustrates the issue of modality redundancy in a daylight scenario. The figure presents the HAR accuracies achieved by various sensor combinations. Although employing the full set of modalities yields the highest accuracy of 96%, a significant insight emerges as the number of modalities is reduced: the drop in sensing accuracy is minimal. For instance, the combinations "RGB+Ra" and "De+Ra" achieve accuracies of 94% and 93%, respectively, only a 2% and 3% decrease compared to the full modality set. This evidence strongly indicates that RGB and depth sensors are redundant, given their overlapping capabilities in detecting target contours and textures.

We can also observe the issue of modality underutilization in a non-LoS setup. As depicted in Fig. 2(b), the accuracy of individual RGB or depth sensors approach the level of random guessing. In stark contrast, radar modality maintains a relatively high accuracy, reaching up to 72%. Interestingly, when RGB and depth sensors are combined with radar, the accuracy does not improve beyond that of the radar alone, remaining around 72%. This lack of improvement suggests that these additional modalities are underutilized. These findings underscore the need for a strategic mechanism to selectively disable both sensors and corresponding computing modules of redundant and underutilized modalities based on specific sensing scenarios, optimizing the effectiveness of the system.

2.2 Time-varying Sensing Environment

In practical scenarios, the sensing environment exhibits time-varying characteristics. For instance, we encounter periodic variations in lighting conditions and electromagnetic fields throughout the day. Additionally, specific events, such

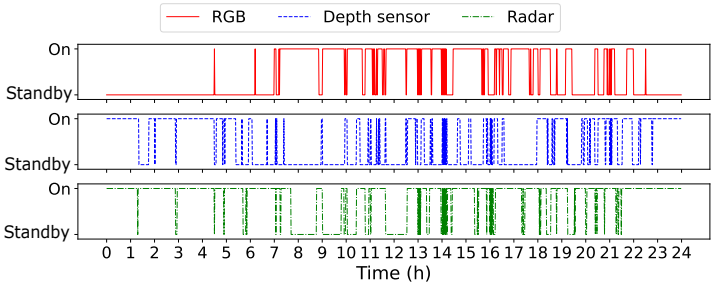


Fig. 3: Optimal modality combination in a day.

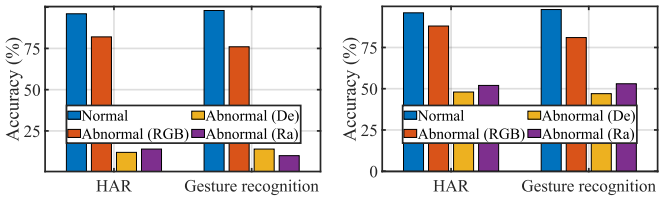
as a human subject entering or exiting a scene, can lead to changes in the environment. In such dynamic contexts, the optimal sensor combinations also evolve. To illustrate this phenomenon, we deploy the sensing system described in § 2.1 to perform Human Activity Recognition (HAR) in a living room over a 24-hour period. We collect full-modality performance data (i.e., accuracy, power usage, and computational overhead) from all sensor and compute modules, and conduct offline traversal experiments with different modality combinations, employing distinct deep learning networks for HAR. Our goal is to identify the optimal sensor setup, i.e., the minimal overhead configuration that achieves an accuracy within $\pm 2\%$ of the full modality's performance. We record the on/standby states of the modalities in the optimal setup in Fig. 3.

One may readily observe the frequent switching between on/standby states of the modalities throughout the day, reflecting that the optimal sensor combination (i.e., the setup with low energy and computational overhead while maintaining sufficient acuracy) is also time-varying. Furthermore, we can observe some long- and short-term trends in Fig. 3. First, there is a clear long-term trend. For example, the RGB camera predominantly enters standby at night while it is activated during the day. In contrast, the depth sensor and radar tend to remain active at night to compensate for the deactivated camera. Secondly, a short-term trend is also noticeable. When the on/standby states of the modalities change, switching oscillations occur within the next few minutes. This phenomenon can be attributed to the dynamic and bursty nature of human activities in the scene. These insights from offline experiments underscore the need for an automatic and real-time modality selection mechanism that effectively addresses both long- and short-term trends in sensing environments.

2.3 Modality Anomaly Due to Deactivation

One issue with deactivating modalities, particularly sensing hardware, is the introduction of abnormal values such as 0's and nulls [34]. This is a crucial problem because the anomalies cannot be handled just by reactivating the sensors. These anomalies can propagate through deep learning computing modules, leading to erroneous results. Although it is feasible to train separate deep learning networks tailored to different modality combinations (as in § 2.1), this approach is impractical due to the substantial computational and storage overhead it necessitates. To illustrate the negative impact of modality anomalies caused by deactivation, we again deploy the sensing system described in § 2.1 for HAR and gesture recognition. Here, anomalous data streams, especially those result from modality deactivation are filled with 0's. The resulting accuracies, with and without anomalous values, are presented in Fig. 4.

One may readily observe that while the baseline accuracies for HAR and gesture recognition exceed 90%, the introduction of abnormal values dramatically decreases performance. For early fusion, which combines data from different modalities at the input stage [35], HAR accuracy falls to 79%, 10%, and 11% when abnormal values are introduced in the RGB camera, depth sensor, and radar modalities, respectively. Similarly, for late fusion, which processes each modality separately before integrating their high-level results [35],



(a) Early fusion. (b) Late fusion.

Fig. 4: Anomalies due to deactivated modalities cause accuracy degradation.

HAR accuracy drops to 85%, 49%, and 52% under the same conditions. Gesture recognition experiences a comparable decline in accuracy. Notably, the RGB modality experiences the least accuracy drop due to it is accustomed to encounter zeros in dark conditions. In contrast, the depth sensor and radar, being active modalities, are relatively unfamiliar with such anomalies. Additionally, late fusion exhibits a smaller accuracy drop compared to early fusion. This is attributed to the reduced coupling among modalities in late fusion, which consequently diminishes anomaly propagation. It is crucial to note that, in practice, these accuracy drops could be even more pronounced due to the time-varying nature of sensing environments, as discussed in § 2.2.

3 System Design

Based on the discussion in § 2, we introduce E-M^2, a framework for efficient multimodal sensing and computing. E-M^2 is described in three main steps, as shown in Fig. 5. First, E-M^2 employs a lightweight DPPN π to dynamically control the on/standby states of the sensing and computing modules. Second, to mitigate the information deficit caused by the deactivated[2] modalities, E-M^2 uses a long-term, short-term temperature scheduling mechanism, strategically activating the deactivated sensing modules, enabling DPPN to acquire complete information and make informed decisions. Third, to combat the accuracy degradation caused by anomalies due to deactivated modalities, E-M^2 incorporates modality-conditioned layers that modulate the features within the sensing network, thus improving its robustness against such anomalies. In the following, we initially concretely define our problem and then delve into the specifics of E-M^2.

3.1 Overview and Problem Statement

The goal of E-M^2 is to achieve efficient multimodal sensing various tasks, including HAR, gesture recognition, and people counting. For a given task, we employ a multimodal system consisting of K input modalities $\{M_1, M_2, \ldots, M_k, \ldots, M_K\}, 1 \leq k \leq K$. Each modality M_k is associated with a specific sensor S_k designed to capture the relevant signals. To optimize the analysis and control of these separate modalities, our system minimizes the coupling between computing modules. Instead of processing data from all modalities simultaneously, E-M^2 utilizes distinct computing modules C_k for each modality, as shown in Fig. 5. Without loss of generality, ResNet-50 is used as the computing module for feature extraction across all modalities [36]. To further enhance the low-coupledness of the modalities, E-M^2 utilizes a late-fusion layer with trainable parameters. This layer fuses the results extracted by each computing module at the final layer, thereby preserving the integrity of the data flow throughout earlier layers.

The key problem E-M^2 is trying to solve is how to selectively turns on or standby both S_k and C_k ("dual-pronged" means controlling both the sensors and computing modules) to achieve optimal energy and computational efficiency. Our approach centers around the development of a DPPN. This network processes data obtained from the sensors S_k and emits control signals to manage the on/standby state of both S_k and C_k, as will be explained in § 3.2. The intertwined nature of sensing and computing modules presents challenges, particularly when deactivating certain sensors. First, deactivating specific sensors can lead to information loss, impeding the DPPN's ability to make accurate decisions. To address this, we propose a method to ensure all sensors are activated at opportune moments, thereby providing comprehensive data input, as will be discussed in § 3.3. Second, entering standby some sensors may result in anomalies for the perception network (i.e., computing modules), as these scenarios are unseen in the training data. To mitigate this issue, E-M^2 informs the computing modules of the current sensor states and modulate their learning processes accordingly. Without loss of generality, we choose RGB camera, depth sensor, and radar modalities to implement E-M^2's prototype, due to their widespread use. Nonetheless, E-M^2 is supposed to generalize to other modalities as well.

3.2 Modality Selection

The challenge of optimal modality selection fundamentally involves making *discrete* decisions: specifically, determining whether sensors should be activated or deactivated. This discrete nature poses a significant obstacle to the application of conventional deep learning techniques, which rely on differentiable processes for optimization through standard backpropagation. One promising approach to address this issue is to frame the decision-making policy within the context of reinforcement learning (RL) [37]. This approach allows for the estimation of policy parameters by calculating the gradient of expected rewards. However, incorporating RL introduces complexities in training, often necessitating variance reduction techniques that add computational overhead. To overcome these challenges, we propose a differentiable dual-pronged policy network (DPPN) π. This network is specifically designed to solve the non-differentiable issue for both hardware and the computing modules, allowing for discrete decision policies aiming at reduced energy as well as computational overheads.

DPPN works as follows: initially, a convolutional feature extractor (with depthwise separable convolutional layers to minimize computational overhead) captures spatial features from the input data. These features are then fed into an LSTM module [38] to further extract temporal causality features. Subsequently, the spatial-temporal features are

2. Throughout this paper, the terms "deactivated" or "disabled" of a E-M^2's sensor refers to a low-power standby mode, not a complete power-off state.

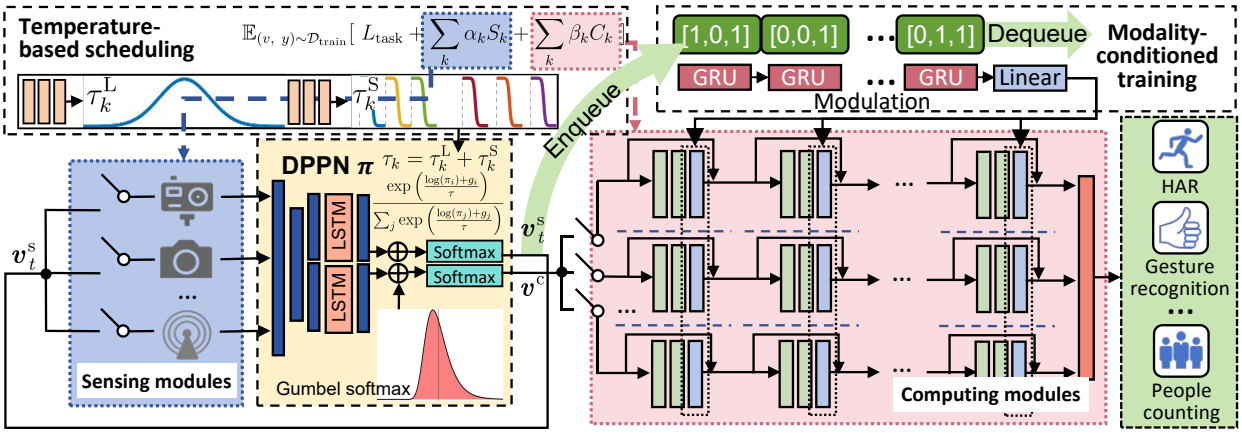


Fig. 5: System design of E-M^2.

mapped to the logit space using a linear layer, yielding scores $\lambda_{k,l}$ for the computing and sensing modules of different modalities $k \in \{1, \ldots, K\}$ at the state $l \in \{\text{on}, \text{standby}\}$. To make discrete decisions $Y_{k,l}$ from the scores, E-M^2 leverages the Gumbel-Softmax [39] technique. Gumbel-Softmax addresses the non-differentiability issue of making discrete decisions by providing a differentiable approximation to the discrete sampling process. By introducing Gumbel noise to the logits and applying the softmax function, Gumbel-Softmax enables gradients to flow through the discrete decisions during training. This facilitates the exploration of different discrete configurations and promotes the learning of robust and generalizable representations, ultimately enhancing the decision-making capabilities of the DPPN.

$$Y_{k,l} = \frac{\exp\left(\left(\log\left(\lambda_{k,l}\right) + g_{k,l}\right)/\tau_k\right)}{\sum_{j\in\{\text{on,standby}\}} \exp\left(\left(\log\left(\lambda_{k,j}\right) + g_{k,j}\right)/\tau_k\right)}, \quad (1)$$

where $g_1, \ldots, g_k$ are i.i.d samples drawn from Gumbel(0,1) [40], τ_k is a temperature parameter [41] for each modality which will be discussed in § 3.3. The dual-pronged outputs of DPPN, denoted as π, simultaneously control the on/standby states of the sensors and the corresponding computing modules. To effectively train DPPN, we define the following loss function, which encourages the selection of modules that minimize power consumption and computational overhead while maintaining sensing performance:

$$\mathbb{E}_{(V,y)\sim\mathcal{D}_{\text{train}}}\left[L_{\text{task}} + \sum_k \alpha_k S_k + \sum_k \beta_k C_k\right], \quad (2)$$

where L_{task} represents standard cross-entropy [42] to measure the classification accuracy, and $\sum_k \alpha_k S_k$, $\sum_k \beta_k C_k$ measure the overhead of the sensing and computing modules for the k-th modality, respectively. In the equation, we have:

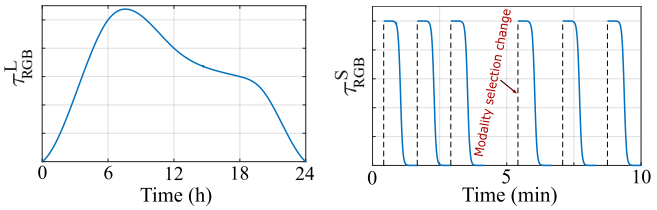
$$S_k = \begin{cases} \left(\frac{|s_k|_0}{p}\right)^2 & \text{correct} \\ \gamma & \text{else} \end{cases}, \text{and } C_k = \begin{cases} \left(\frac{|c_k|_0}{p}\right)^2 & \text{correct} \\ \xi & \text{else} \end{cases}, \quad (3)$$

where $S_k = \left(\frac{|s_k|_0}{p}\right)^2$ and $C_k = \left(\frac{|c_k|_0}{p}\right)^2$ represent the proportion of time the k-th sensor and the corresponding computing module is in the "on" state relative to the entire time span p, assuming correct predictions during the training phase. Additionally, $S_k = \gamma$ and $C_k = \xi$ denote the fixed cost associated with prediction errors, which acts as an incentive for the model to minimize classification errors.

3.3 Long Short-term Temperature Scheduling

In § 2.2, we have observed that sensing environments exhibit time-varying characteristics, leading to the evolution of the optimal sensor combination over time. Consequently, it is impractical to perform modality selection just once and rely on it indefinitely. Although the DPPN, as discussed in § 3.2, can automatically select modalities, it lacks sufficient exploration capabilities and may become stuck in suboptimal configurations. Unlike sensors-always-on systems, this issue is particularly exacerbated when sensors enter standby: DPPN cannot obtain sufficient information from an inactive sensor, which may consequently never be reactivated. To address this limitation, one potential mechanism for enhancing exploration is by increasing the temperature τ in Eq. (1). A higher temperature results in a more uniform logit distribution, thereby encouraging the DPPN to explore different on/standby states for sensors that have already entered standby. Conversely, a lower temperature promotes exploitation by increasing the discrepancy in the logit distribution, making the system more prone to make the "most likely" decisions. To strike a balance between exploitation and exploration that achieves maximal efficiency, we should come up with a temperature scheduling mechanism that adapts to the environment.

To effectively manage temperature scheduling, it is crucial to understand the two trends in Fig. 3: the long-term trend characterized by daily fluctuations, and the short-term trend that follows in modality state changes. The long-term temperature τ_k^{L} predictably shifts over time in response to circadian cycles influenced by ambient light variations, environment dynamics, and human activities. In contrast, the short-term temperature τ_k^{S} plays a vital role in adapting to abrupt modality changes that may result in incomplete modal data. An elevated τ_k^{S} becomes crucial under such incomplete conditions, as it reactivates the sensor, thereby empowering DPPN with the necessary data to accurately assess environment stability. Furthermore, if the system identifies multiple environment oscillations within a set time period (practically set to 1 min), τ_k^{S} will not reset,



(a) Long-term temperature. (b) Short-term temperature.

Fig. 6: Long- and short-term temperatures.

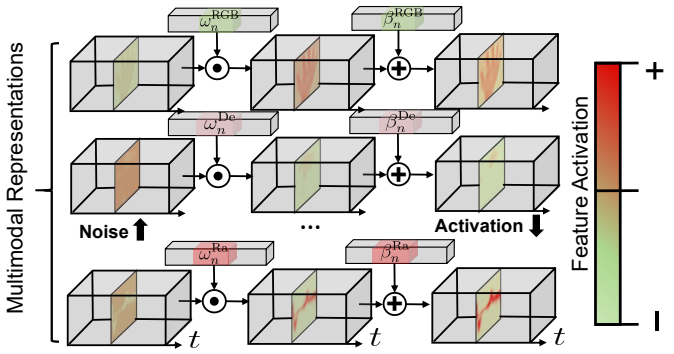


Fig. 7: E-M^2 modulates temporal multimodal representations by modality-conditioned training.

leading to a continuous reduction in temperature aimed at preventing potential instability. Thus, the overall temperature is expressed as $\tau_k = \tau_k^{\mathrm{L}} + \tau_k^{\mathrm{S}}$, encompassing both long- and short-term trends.

In the following, we detail the implementation of τ_k^{L} and τ_k^{S}. A simplistic method might involve using fixed temperatures for every sensing environment. However, given the varied nature of these environments and their associated tasks, it is crucial for E-M^2 to adapt rather than apply a one-size-fits-all solution. Therefore, we employ a learning-based strategy, leveraging multilayer perceptrons (MLPs) [43] that are designed to learn and adjust the temperatures dynamically. Specifically, and $\tau_k^{\mathrm{L}} = \mathrm{MLP}_k^{\mathrm{L}}(t)$ and $\tau_k^{\mathrm{S}} = \mathrm{MLP}_k^{\mathrm{S}}(t)$, where t represents the current timestamp. These MLPs are trained concurrently with DPPN, using the timestamp as input to output the appropriate temperature for that moment. This mapping can be effectively learned because the MLPs receive a rich, task-relevant gradient from the subsequent computation graph, and the underlying relationship between the scalar time input and the scalar temperature output is highly structured. We present an example of the trained $\tau^{\mathrm{L}}_{\mathrm{RGB}}$ and $\tau^{\mathrm{S}}_{\mathrm{RGB}}$ in a living room inhabited by a family of three. The results shown in Fig. 6 reveal that the long-term temperature $\tau^{\mathrm{L}}_{\mathrm{RGB}}$ is elevated during the day and decreases at night, aligning with the family's daily routine. Furthermore, $\tau^{\mathrm{S}}_{\mathrm{RGB}}$ starts at a higher value, which aids in reactivating the RGB camera to gather comprehensive environment data, before gradually declining as the environment becomes stable. Note that the temperatures are designed to be a "good fit" that renders robust temporal priors but do not provide the system's ultimate generalizability; the results shown in Fig. 6 are illustrative examples of this mechanism's behavior. In practice, our system learns aggregate temperature accounting for all scenarios.

3.4 Modality-conditioned Training

Although our design in § 3.3 alleviates the negative impact of modality deactivation, the negative impacts of anomalous values makes it insufficient to only schedule the temperatures. As stated in § 2.3, deactivated sensors during system operation might introduce anomalies unseen during training, thus negatively impacting sensing performance. This discrepancy arises not only because of feature missing, but also because traditional training assumes full-modality activation, an assumption that breaks down in scenarios where both sensors and computing modules are dynamically controlled. When some sensors are deactivated, data distribution shifts occurs, leading to a mismatch between the training and operational environments. To address this issue, we should come up with a method to explicitly inform the computing modules of the on/standby states of the sensors, therefore guiding a focused analysis on the sensing data. To achieve this, we employ modality-conditioned training to condition the computing modules on the sensor states, allowing it to focus more on effective information while dropping out anomalies. This approach ensures that the system can dynamically adapt to changes in sensor states, maintaining robust performance even in the presence of unexpected anomalies.

We present the design of the modality-conditioned training in Fig. 5. The module incorporates a modality-conditioned queue $\mathcal{Q}$ of length p to store history of sensor switching decisions. When the DPPN makes a new binary-encoded sensor-related switching decision $\boldsymbol{v}_t^{\mathrm{s}}$ to control the sensors, $\boldsymbol{v}_t^{\mathrm{s}}$ is simultaneously enqueued in temporal order, dequeuing the oldest control vector at the front of the queue. Consequently, at time t, $\mathcal{Q} = \left[\boldsymbol{v}_{t-p+1}^{\mathrm{s}}, \boldsymbol{v}_{t-p}^{\mathrm{s}}, \cdots, \boldsymbol{v}_t^{\mathrm{s}}\right]$. The contents of $\mathcal{Q}$ are then input into a data fusion module consisting of Gated Recurrent Units (GRUs) [44] for extracting temporal features from the sensor switching decision history. Subsequently, a linear layer fuses these features and produces weights and biases (ω_n^k, β_n^k) to modulate the n-th temporal representation of the k-th modality as follows:

$$F(\mathbf{R}_n^k | \omega_n^k, \beta_n^k) = \omega_n^k \mathbf{R}_n^k + \beta_n^k, \tag{4}$$

where $\mathbf{R}_n^k$ represent the n-th temporal representation of the k-th modality. The proposed affine transformation effectively modulates $\mathbf{R}_n^k$. This modulation process is achieved through a Hadamard product [45] between the temporal representations and learnable weights ω_n^k, followed by the addition of learnable biases β_n^k. This process enhances the feature strength of valid information, reduces the focus on invalid parts. Meanwhile, the biases β_n^k help fill in valid values and prevent anomalies, further enhancing the quality of the modulated representations. The modality-conditioned training strategy is computationally efficient, aligning with E-M^2's overall efficiency goal, as it employs only an affine transformation. Fig. 7 illustrates the modulation process in the context of a gesture recognition task, showcasing how the temporal representations from RGB and depth sensors (capturing the human hand) and the spectrogram from radar sensors are weighted and biased.

4 Implementation

In this section, we describe the dataset, implementation, and the evaluation setup used by E-M^2.

4.1 Dataset

Due to the lack of publicly available datasets suited to our needs, we have collected our own dataset covering sensing tasks including HAR, gesture recognition, and people counting. The entire dataset is not used across all tasks. Instead, three independent datasets are collected, each dedicated to training and evaluating a corresponding model for an individual task. For HAR, we have captured data of 12 common human activities: walking, running, jumping, squatting, turning around, sitting down, standing up, falling down, fencing, bending, nodding, and leg raising. For gesture recognition, we focus on 10 specific gestures: swiping left, swiping right, swiping up, swiping down, pinching in, pinching out, tapping, double tapping, and rotating both clockwise and counterclockwise. These gestures are chosen for their relevance and practical application in user interfaces. The people counting task is designed to accurately discern the number of individuals present in a room, offering valuable insights for space management and monitoring.

Our data covers various sensing orientations across 20 distinct scenarios: a auditorium, a hallway, a living room, a kitchen, a bathroom, a dining room, a balcony, a library, a gymnasium, 2 meeting rooms, 3 offices, 3 classrooms and 3 bedrooms. Four extra sets are collected for further evaluations, which is detailed described in §5.2. For each class in each scenario in HAR and gesture recognition tasks, we collect over 20,000 data samples by RGB camera, depth sensor and radar. To synchronize the modalities, we perform a rapid hand-waving in front of the sensors before the subjects starts their actions. Subsequent data from each modality is aligned according to the timing of this motion. For people counting tasks, each scenario comprises 0-20 individuals in unique positions and poses. Over 20,000 data samples are collected per scenario. For each of the three task-specific datasets, the data is partitioned into a 80% training set and a 20% test set. A 10% validation set is randomly sampled from training set for each fold of the cross-validation. This process ensures all partitions contain samples from all scenarios. The experiments strictly follow the IRB of our institution.

4.2 System Implementation

Hardware implementation: E-M^2 employs the following sensors: a XeThru X4M05 IR-UWB radar [46] with a frequency range of 7.29∼8.75 GHz and a maximum sampling rate of 2042 fps/s; a DCAM710 depth sensor [47] with a detection range from 0.35 m to 6 m, resolution of 640×480 and maximum frame rate of 30fps; and an RGB camera with a resolution of 1920×1080 and frame rate of 30fps. To ensure the consistency across the modalities, we place them on a fixed tripod horizontally as shown in Fig. 8. A GPU workstation equipped with two NVIDIA GeForce RTX 4090 [48] graphics cards is used for training E-M^2. We deploy E-M^2 on a NVIDIA Jetson™ TX2 [49] edge computing device during the inference stage. We also employ the WASITES PZ9002 digital power meter [50] with a power resolution of 0.01 W, to realize precise power consumption measurements.

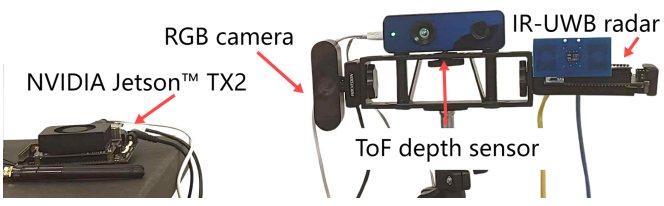


Fig. 8: E-M^2 prototype.

Software implementation: Python 3.11.4 and PyTorch 2.2.1 with CUDA 12.1 support are used for developing the software of E-M^2. The computational overhead of the system, measured in Giga Floating-point Operations (GFLOPs), is evaluated using NVIDIA Nsight tools [51]. We use the software FFmpeg [52] to perform data resampling and standardization for aligning different modalities and easier processing by the neural networks.

Power state transition protocols: The state switching is managed by the NVIDIA Jetson TX2 host. For RGB camera, the host sends suspend commands via USB 2.0 protocol using standard UVC power management. The USB controller halts video streaming, powers down the CMOS image sensor and analog front-end circuits, while maintaining USB interface logic for bus enumeration and wake-up detection. For the depth sensor, the host sends standby commands via USB 2.0 using Vzense SDK. The internal controller then halts data streams and powers down IR VCSEL projector and image sensors, maintaining USB interface circuitry for resume detection and power monitoring logic. For radar, the Jetson host commands the X4M05 module via USB 2.0 protocol to enter standby. The module's controller ensures data acquisition completion, then configures X4 chip power registers via internal SPI to power down UWB transceiver and clocks while maintaining embedded controller, low-power oscillator, and wake-up logic.

4.3 Evaluation Setup

To evaluate E-M^2, we define the evaluation metrics including power consumption, computational overhead, and an additional metric for sensing performance. In the context of HAR and gesture recognition, this performance metric is classification accuracy. For the people counting task, we approach it as a classification problem, defining the metric as the accuracy in predicting the number of individuals.

5 Evaluation

In this section, we evaluate the performance of E-M^2 under various experiment setups mentioned in 4.3. We also discuss the generalization capabilities, analyze the impact of practical factors, and perform analysis of E-M^2's modules.

5.1 Overall Performance

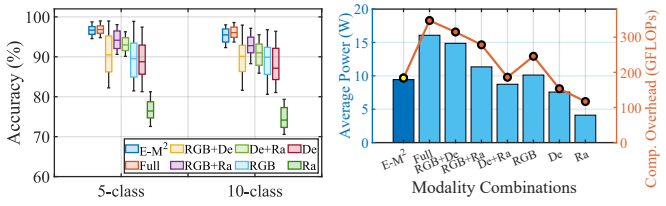
In this subsection, we compare the performance of E-M^2 with the baseline method for three tasks: HAR, gesture recognition, and people counting. As a reference, we record

the average power consumption and computational overhead of the RGB camera, depth sensor, and radar in running mode/standby mode during a single test. The power consumption of RGB camera, depth sensor, and radar are 6.03 W, 4.37 W, and 1.24 W in running mode, and 0.22 W, 0.20 W, and 0.08 W in standby mode, respectively. Moreover, the computational consumption of RGB camera, depth sensor, and radar in running mode are 161.45 GFLOPs, 69.13 GFLOPs, and 32.67 GFLOPs, and all zeros in standby mode, respectively.

5.1.1 HAR

Fig. 9 provides a comparative analysis of E-M^2 and conventional baseline methods in the context of the HAR task. As depicted in Fig. 9(a), E-M^2 demonstrates robust accuracy in the HAR task. The figure reveals that irrespective of whether the problem involves 5 classes or 10 classes, the average accuracy of E-M^2 is on par with methods that leverage all modalities simultaneously, and it significantly outperforms any dual-modality or single-modality approaches. E-M^2 demonstrates enhanced stability in performance when classes of the task increases for its capability to adapt to various situations. Unlike the baseline methods, which experience a notable drop in accuracy from the 5-class to the 10-class problem, E-M^2 maintains its performance and even surpasses the full-modality baseline method in certain metrics. This can be potentially attributed to its capability to selectively deactivate modalities that are detrimental and potentially distracts the current task.

Fig. 9(b) reveals that E-M^2 reduces system power consumption by approximately 41.35% compared to the baseline method with full modalities, including the power consumption of edge devices. Its energy efficiency even exceeds that of some dual-modality combinations. The computational overhead curve exhibits a similar pattern to power consumption. Overall, E-M^2 reduces computation by about 47.12%, which is only 83.22% of the RGB+De combination. The single-modality data indicate that the RGB camera is the primary contributor to power consumption and computational overhead. E-M^2 effectively mitigates these demands through on-demand control of the RGB modality.

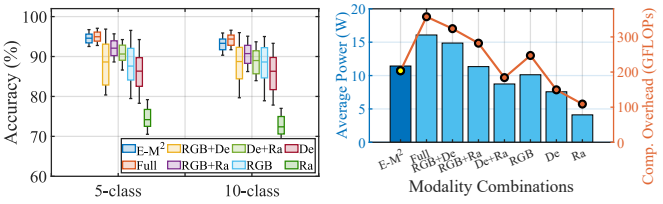


(a) Accuracy. (b) Overhead.

Fig. 9: Performance of HAR.

5.1.2 Gesture Recognition

Fig. 10(a) reports the results of comparing E-M^2 with baseline methods for the gesture recognition task. Due to the smaller amplitude of movements, gesture recognition tasks are harder to distinguish compared to HAR, resulting in a certain drop in overall accuracy. Nevertheless, the baseline method exhibits a reduction in accuracy ranging from 3.13% to 9.47%, primarily attributed to the inherent challenges radar faces in distinctly differentiating subtle hand movements, while E-M^2 experiences a marginal decline of approximately 2.7%. This once again proves the previously discussed robustness of E-M^2.



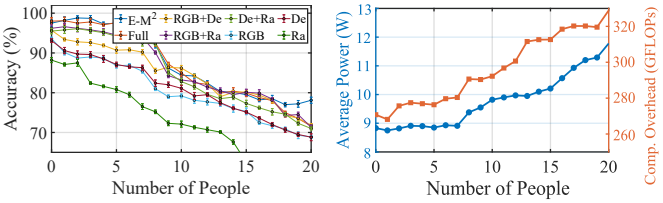
(a) Accuracy. (b) Overhead.

Fig. 10: Performance of gesture recognition.

As shown in Fig. 10(b), the comparative analysis of E-M^2 and baseline methods regarding power consumption and computational overhead reveals that the gesture recognition and HAR tasks exhibit similar trends. E-M^2's power consumption and computational overhead are situated between those of most dual-modality combinations and single-modality ones. A notable phenomenon is the significant increase in power computational overhead for gesture recognition when compared with HAR, possibly attributed to the necessity of capturing subtle hand movements which demands higher spatial resolution for detailed information extraction. Despite the overhead increase, E-M^2 strives to decrease these overheads by strategically disabling redundant modalities, underscoring its superior efficiency.

5.1.3 People Counting

The people-counting accuracy of E-M^2 is shown in Fig. 11. As seen in Fig. 11(a), the overall sensing capability of E-M^2 closely matches that of its full-modality counterpart while outperforming all other modality combinations. Notably, the counting accuracy decreases as the number of people increases from 0 to 20. This is likely because as the scene becomes more crowded, the probability of both inter-person occlusion and partial person-object occlusion rises, making it more challenging to distinguish individuals. Each modality combination presents a unique accuracy curve with a distinct "turning point," after which the rate of decline becomes more pronounced. This observation can potentially be attributed to the limitations in the sensor's resolution. Despite this decline, E-M^2 demonstrates a more stable accuracy as the number of people increases when compared to the baselines.



(a) Accuracy. (b) Overhead.

Fig. 11: Performance of people counting.

We also inspect the power and computational overhead as a function of the number of people, as shown in the Fig. 11(b). It is shown that both power consumption and computational overhead increase progressively with the number of participants, which is particularly pronounced when the number exceeds 7. This behavior highlights the E-M^2's intelligent adaptation to scenarios with more participants, as it strategically allocates extra resources to maintain sensing performance in more complex and crowded scenes. Additionally, we note two key observations from our broader analysis that are not directly shown in this figure: first, on average, E-M^2 reduces power and computational overhead by 39.5% and 35.6% respectively compared to the full-modality baseline; second, E-M^2's power and computational requirements slightly exceed those of a single depth sensor. This demonstrates an effective decision, bypassing the power-intensive RGB camera to strategically leverage the more efficient depth sensor.

5.2 Generalization Capabilities

In this section, we evaluate the generalization capability of E-M^2. The scenarios are notably diverse, featuring various environments and human subjects with differing heights, weights, postures, and other physical traits.

5.2.1 Generalization to Different Environments

We further evaluate the generalization capability of E-M^2 in 4 environments (lab, bedroom, office and playground) unseen during training. Without loss of generality, we fix the number of people in the scene to 5 for the people counting task. The results of our experiments are illustrated in Fig. 12(a). It is revealed that for the HAR task, the sensing accuracy is minimally impacted, experiencing a decline of no more than 2.87%. Although the gesture recognition and people counting tasks face slight challenges due to variations in hand poses and individuals' positions and postures, E-M^2 consistently performs with an accuracy exceeding 90%, even in the most complex situations. This demonstrates E-M^2's robust performance and adaptability in unfamiliar environments.

We further inspect the system's power consumption and computational overdhead when generalizing to unseen environments in Fig. 12(b). For all three tasks, different environments cause significant variations in the system's modality usage decisions, leading to substantial fluctuations in power and computational overhead. Nevertheless, for all tasks, E-M^2's consumption remains under 70% of the full modality as mentioned in § 5.1, demonstrating E-M^2's capability to achieve efficiency by deactivating unnecessary sensors and their corresponding computing modules in complicated environments.

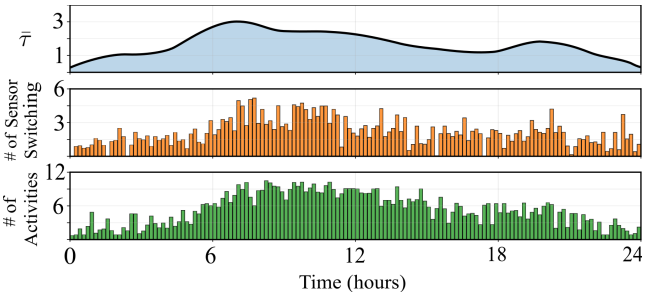
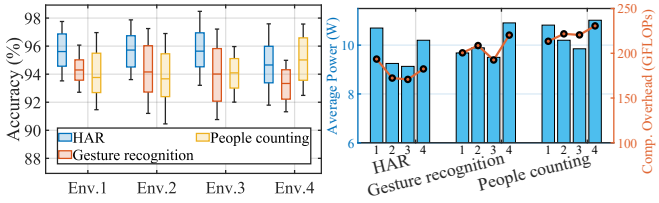


Fig. 13: τ obtained by the long short-term temperature scheduling effectively captures the aggregate activity dynamics across various scenarios.

To provide evidence of the system's generalizability and the effective fit of the long short-term temperature scheduling mechanism, we sample 7 days of data across all scenarios and plot the average daily trends of: i) $\bar{\tau} = \bar{\tau}^{\mathrm{L}} + \bar{\tau}^{\mathrm{S}}$ (aggregated from all scenarios and averaged across all sensors and days), ii) the number of sensor state changes and iii) the number of human activities. The results visualized in Fig. 13 illustrates an excellent fit among the three trends. It is shown that higher values of $\bar{\tau}$ correspond to periods of more frequent human activities, which in turn leads to a higher rate of sensor state switchings made by the DPPN; while lower $\bar{\tau}$ values align with the stable periods. This synergy between the universal temporal prior and the data-driven policy is the cornerstone of E-M^2's ability to generalize effectively to new environments.

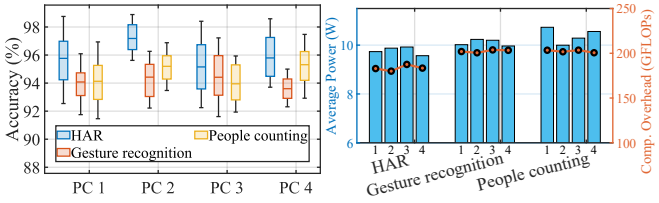
5.2.2 Generalization to Different Human Subjects

We further investigate the generalization capabilities of E-M^2 across different human subjects, presenting the results in Fig. 14. To maintain consistent notations across various tasks, we use "PC" to denote people combinations. In the context of HAR and gesture recognition, each human subject is considered a single "PC," while in people counting, a group of people with the same number constitutes one "PC." Fig. 14(a) demonstrates that variations in people combinations have a significant impact on HAR outcomes due to the differing movement habits and patterns among human subjects. In contrast, for gesture recognition, the impact is less pronounced, likely because subjects are asked to perform standardized gestures. For the people counting task, sensing performance discrepancies across different people combinations are minimal. This is likely because the



(a) Accuracy. (b) Overhead.

Fig. 12: Generalization to different environments.



(a) Accuracy. (b) Overhead.

Fig. 14: Generalization to different human subjects.

neural network prioritizes the group as a whole rather than focusing on individual habits.

We demonstrate how power consumption and computational overhead are affected when adapting to new human subjects. As shown in Fig. 14(b), E-M^2 maintains consistent overhead across various combinations of individuals, a trend in stark contrast with the outcomes in § 5.2.1. This consistency suggests that variations in human subjects do not substantially alter the data features influencing the decision-making of the DPPN. However, significant variations in overhead are observed in the people counting task. This can likely be attributed to the fact that the presence of multiple individuals alters the environment, thereby impacting the DPPN's decision-making process.

5.3 Impact of Practical Factors

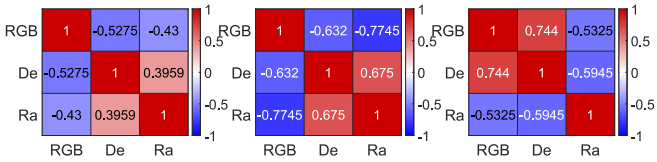
We study the impact of practical factors in this section. Note that all the HAR and gesture recognition tasks are still evaluated in a single-person context, aligning with the scope of our work. For the people counting task, the distributions of the number of participants are consistent across different factors. Unless otherwise noted, all evaluation metrics reported in this section represent the average performance across the three tasks.

5.3.1 *Illumination*

We conduct experiments on the system under typical indoor and outdoor lighting conditions (0 to 20000 lux). As the results shown in Fig. 15, the system maintains an accuracy of over 89% under extremely low light conditions ($\leq$ 10lux). This is attributed to the use of depth sensors and radar modalities, which are unaffected by lighting conditions. Subsequently, the system opt to activate the RGB camera modality, achieving an accuracy of over 95%. The power consumption curve shows a corresponding increase, further validating the system's decision. Thereafter, as the illumination increases to over 19000 lux, the RGB images become overexposed, prompting the system to deactivate the RGB camera once again.

5.3.2 *Time of day*

Fig. 16 reveals the impact of the time of day. Throughout the 24-hour cycle in one day, various factors such as illumination, human activities, and electromagnetic field undergo continuous changes. These fluctuations result in deviations of the system's accuracy and workload. As an example, the system demands more energy and computation during the day (7:00 to 18:00) than at night (18:00 to 7:00) because it needs to process more frequent human activities. However, despite these external influences, E-M^2 consistently maintains an accuracy above 90%, demonstrating exceptional robustness. Additionally, although certain special circumstances necessitate higher power consumption to achieve sensing accuracy, the average power consumption remains within 9 W, ensures excellent energy efficiency while maintaining high accuracy.



(a) HAR. (b) Gesture recognition. (c) People counting.

Fig. 17: Correlation matrices of sensor states.

Over a 24-hour monitoring period, we observe a interesting phenomenon where certain modalities shows significant correlations. This prompts us to investigate potential improvements. We plot the correlation matrices of all modalities in different tasks in Fig. 17. According to Fig. 17(a), RGB cameras and depth sensors exhibits a marked negative correlation in HAR tasks, likely due to their alternating operation with the 24-hour cycle. In gesture recognition tasks, as illustrated in Fig. 17(b), depth sensors and radar show a strong positive correlation, because both depth senor and radar have strong capabilities to accurately capture depth information, while RGB camera can only capture information such as color and contour. Considering the significance of depth in gesture recognition, it is plausible to assume that RGB camera might remain inactive for longer durations in this task. In people counting tasks, as depicted in Fig. 17(c), a strong correlation between RGB and depth sensors is noted, suggesting that radar's resolution might be inadequate for densely populated scenes, whereas the other two sensors are more effective in such scenarios.

5.3.3 *Occlusion*

Target occlusion is a critical factor influencing perception accuracy. We define occlusion as the obstruction of the view of sensors. In our experiments, we access its impact by measuring the system performance while changing the sensors' Field of View percentage (%FoV). Results in Fig. 18 indicated that when %FoV is below 42%, RGB cameras and depth sensors are nearly ineffective, leading E-M^2 to deactivate these modalities for better power and computation efficiency. Beyond this threshold, the system activates all modalities to enhance accuracy. After 85%, RGB sensors alone suffice for accurate sensing, allowing the system to deactivate redundant radar and depth modalities to for optimized efficiency.

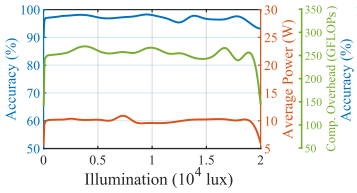


Fig. 15: Illumination.

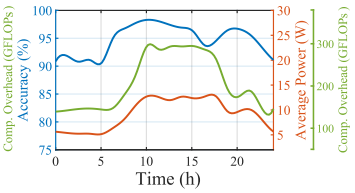


Fig. 16: Time of day.

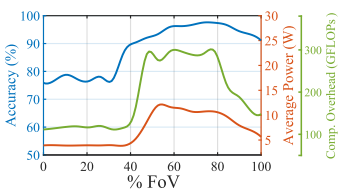


Fig. 18: Occlusion.

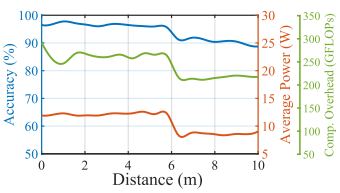


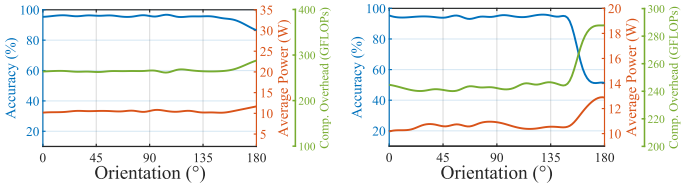
Fig. 19: Sensing distance.

5.3.4 Distance

The distance between the sensor and the sensing target also markedly influences accuracy. System performance in different sensing distances is shown in Fig. 19. Within the sensor's effective range, sensing accuracy declines slowly as the distance increases. A minor accuracy drop shows up at around 6.0 m, which is the maximum operational distance of the depth sensor used in this study. At greater distances, accuracy experiences a brief improvement, along with increased power consumption and computational overhead, suggesting the system engages additional modalities to maintain sensing functionality. Within the experimental range of up to 10 m, the worst accuracy observed is 89.75%, which is adequate for most indoor and outdoor sensing applications.

5.3.5 Participant Orientation

Beyond the factors previously discussed, we further investigate the impact of participant orientation, focusing on HAR and gesture recognition tasks as people counting is fundamentally insensitive to facing directions. The results are shown in Fig. 20. For the HAR task shown in Fig. 20(a), the system maintains stable, with accuracy consistently exceeding 95% across a wide orientation range from 0° to 150°. While a slight decline is observed as the participant reaches a full 180° back-facing orientation, the system actively preserves high performance by allocating more resource. Gesture recognition is more sensitive to orientation as Fig. 20(b) reveals. The system's performance is stable up to 145°, after which it gradually decreases to a final accuracy of approximately 52% at 180°. Notably, accuracy remains significantly above random chance even when the torso completely obstructs gestures beyond 145°. This suggests that E-M^2 effectively utilizes radar and learns to infer from secondary cues, such as subtle torso movements. Despite this performance variation, the increase in system overhead at extreme angles is similar for both HAR and gesture recognition. This reveals an adaptive mechanism where E-M^2 strategically trades higher power and computational overhead to preserve sensing performance under challenging conditions.



(a) HAR. (b) Gesture recognition.

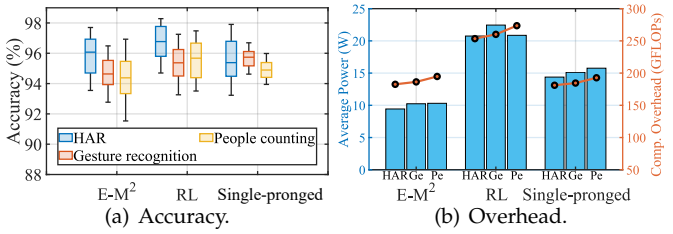
Fig. 20: System performance on various orientations.

5.4 Module Analysis

In this section, we analyze the effects of E-M^2's modules.

5.4.1 DPPN

DPPN functions to decide the on/standby states of sensors and their corresponding computing modules based on sensing environment. Besides DPPN, RL methods [53] and single-pronged approaches [28], [54] can also achieve similar goals. However, not all methods are equally effective. Fig. 21 presents a comparison among them. The RL method achieves the highest accuracy across three tasks. However, its design necessitates sophisticated strategy optimization and value function estimation, alongside the management of high-dimensional state spaces during environment interactions, leading to significant power consumption and computational burden. Therefore, despite the RL method's superior accuracy, E-M^2 is more appropriate for practical edge environments due to its manageable overhead.



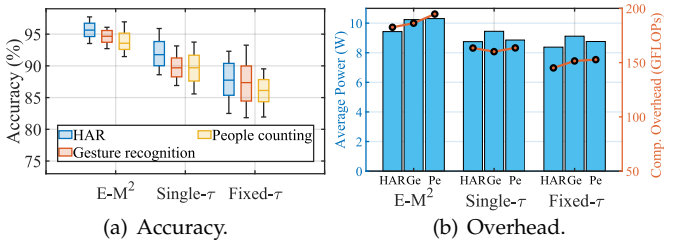
(a) Accuracy. (b) Overhead.

Fig. 21: Comparison of DPPN module and its counterparts on system performance.

On the other hand, the single-pronged method, which perpetually activates all sensors to acquire environment data and make comprehensive decisions on whether to activate the corresponding computing modules, achieves high accuracy by relying on complete environment information. This approach can significantly reduce computational overhead but still fails to address the problem of excessive system power consumption according to Fig. 21(b). Keeping all sensors active at all times means the system consumes energy even when not needed, resulting in low overall energy efficiency. DPPN has comprehensive and unique advantages over the other two methods. It integrates the high accuracy of RL methods with the low computational cost of single-pronged approaches, making it a more ideal choice in edge computing environments.

5.4.2 Long Short-term Temperature Scheduling

E-M^2 is trained with a dynamic τ that accounts for both long-term and short-term variations in the sensing environment. For comparison, we assess E-M^2 against two alternative approaches: one involves training a single τ, and another with a fixed τ. The comparative results are illustrated in Fig. 22. Specifically, Fig. 22(a) highlights the accuracy outcomes, demonstrating that E-M^2 consistently



(a) Accuracy. (b) Overhead.

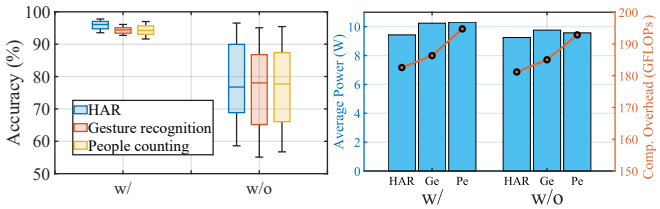
Fig. 22: Comparison of long short-term temperature scheduling on system performance.

surpasses the other methods by approximately 3% and 7% across all tasks. This performance quantitatively validates the effectiveness of E-M^2 in learning the dynamics of τ, indicating that employing long short-term temperature scheduling effectively captures both daily and short-term environment dynamics. Conversely, the single-τ method struggles to adapt to changes, often leading to improper modality usage strategies that overlook critical information. This limitation is even more pronounced when the temperature is fixed.

Fig. 22(b) further presents a comparative analysis of energy and computational overhead for E-M^2, the single-τ method, and the fixed-τ approach. Throughout a typical sensor monitoring cycle, E-M^2 demonstrates superior sensing performance through dynamic adaptation to changing environments. This enhanced performance comes with the a slight increase in power and computational overhead: compared with the single-τ method, E-M^2 incurs an additional overhead of 1 W and 30 GFLOPs. When compared with the fixed-τ method, these numbers rise to approximately 2 W and 50 GFLOPs. Despite these slight increases, the improvement in sensing performance is substantial, underscoring the efficacy of E-M^2.

5.4.3 *Modality-conditioned Training*

We conduct an ablation study to evaluate the impact of the modality-conditioned training strategy of E-M^2. Fig. 23 presents a performance comparison between models trained with and without this strategy, revealing an average accuracy improvement of 18.62% across various tasks when employing the modality-conditioned training strategy. While this strategy significantly enhances accuracy, it does introduce some extra overhead. This is primarily due to the maintenance of the queue within the modality state pipeline and the integration and synthesis of data streams within the GRUs in the data fusion pipeline. However, this overhead is minor compared to the scale of the sensing neural network, rendering it negligible. In conclusion, the improvements in both accuracy and efficiency make this strategy a worthwhile enhancement to multimodal sensing and computing systems.

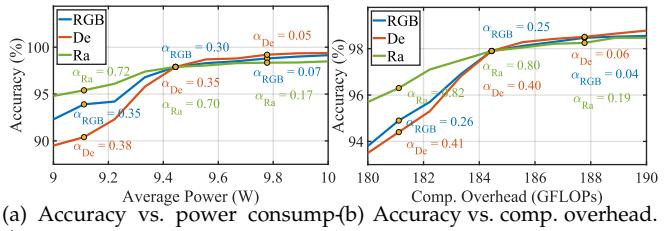


(a) Accuracy. (b) Overhead.

Fig. 23: Sensing performance w/ and w/o modality-conditioned training.

5.5 Hyper-parameter Tuning

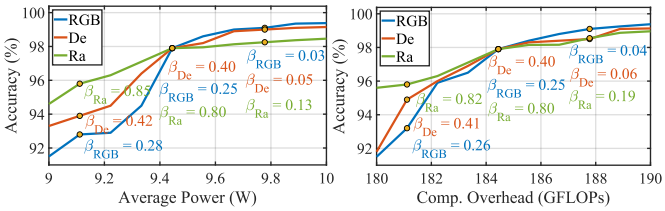
We examine how hyper-parameters influence the performance of E-M^2, specifically focusing on the weights α_k and β_k in the loss function described in Eq. (2). Without loss of generality, we choose to evaluate on the HAR task.



(a) Accuracy vs. power consump-tion. (b) Accuracy vs. comp. overhead.

Fig. 24: Impact of α on accuracy and overhead.

During our analysis, we systematically adjust one hyper-parameter at a time while keeping all others at their default settings: $\alpha_{\text{RGB}} = 0.3$, $\alpha_{\text{De}} = 0.35$, $\alpha_{\text{Ra}} = 0.75$, $\beta_{\text{RGB}} = 0.25$, $\beta_{\text{De}} = 0.4$, and $\beta_{\text{Ra}} = 0.8$. The default values ensure E-M^2 maintains a sensing accuracy within 1% of its full-modality version. Figures 24 illustrates the effects of varying α_k on sensing accuracy and overhead. Our analysis reveals a general trend: increased accuracy coincides with higher power consumption or computational overhead. This observation aligns with information theory principles, where an enhanced signal-to-noise ratio allows for better information extraction about the target. Furthermore, increasing α_k acts as a penalty for excessive power and computation , reducing both overheads. Notably, tuning α_{De} significantly impacts accuracy, because the depth sensor offers substantial information with minimal power and computational demands.



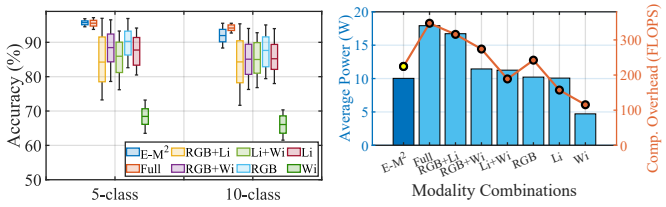
(a) Accuracy vs. power consump-tion. (b) Accuracy vs. comp. overhead.

Fig. 25: Impact of β on accuracy and overhead.

Fig. 25 clearly demonstrates the impact of adjusting the overhead weight β_k associated with computing modules. One may observe a positive correlation persists between accuracy and average power/computational overhead, mirroring the effects observed when fine-tuning α_k. However, unlike α_k, β_k influences only the computing module usage, resulting in a comparatively smaller effect on overall system performance. Notably, sensing accuracy is most sensitive to β_{RGB} linked with the RGB camera, owing to the rich information it gathers and the feature extraction by computing modules. Together, Figures 24 and 25 offer valuable insights into selecting optimal parameters α_k and β_k, thereby enabling adaptations to diverse power and computational constraints in varying edge environments.

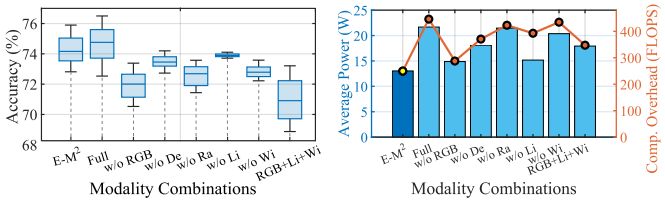
5.6 Modality Extension

In this section, we conduct experiments with additional modalities to validate the extensibility of our system. We incorporate Livox Mid-360 LiDAR [55] (denoted as "Li") and the 802.11ax channel state information (CSI) of the



(a) Accuracy. (b) Overhead.

Fig. 26: Extending E-M^2 to more modalities.



(a) Accuracy. (b) Overhead.

Fig. 27: HAR performance of E-M^2 on the MM-Fi dataset.

ASUS RT-AX86U Pro router [56] (denoted as "Wi") to perform HAR. These two modalities are particularly valuable for HAR applications: LiDAR captures pointclouds with contour and reflectivity information of the target, while WiFi detects human activity by channel variations caused by interactions between RF signals and the target. However, the computational resources and hardware constraints in edge environments, such as insufficient data ports mentioned in § 1, make it challenging to achieve synchronization and efficient data processing across all modalities on a single computing unit. As a practical alternative, we implement a 3-modality configuration that combines RGB camera, LiDAR, and WiFi for supplementary experiments.

The results depicted in Fig. 26 demonstrate the overall performance of the corresponding models. Comparing the 7-th box of Fig. 26(a) and Fig. 9(a), LiDAR demonstrates superior accuracy compared to the depth sensor, mainly because it yields richer contour and geometric information for the recognition network. Similarly, the 8-th box in both figures reveals that WiFi exhibits lower accuracy than radar. This discrepancy stems from WiFi signals' inherently lower range and angular resolution. Moreover, the communication-oriented beam-forming technique further diminishes the signal quantity available for sensing purposes. Despite these, E-M^2 strategically utilized the information from WiFi together with other modalities, enabling the system to achieve an 95.7% mean accuracy. Fig. 26(b) shows the participation of LiDAR introduces additional energy and computational costs. Nevertheless, when compared with full modalities, E-M^2 demonstrates 44.03% energy savings and 35.38% reduced computational overhead while sustaining high accuracy performance, aligning with our previous experimental results.

5.7 Extension to Other Datasets

In this subsection, we introduce that E-M^2 functionally has the ability to support more modalities and activities. We extend E-M^2 to MM-Fi dataset [57] for offline simulation experiments. This HAR dataset features a comprehensive collection of five key modalities: RGB images, depth images, LiDAR point clouds, radar data, and WiFi CSI data. This robust combination captures 27 distinct human activities, all thoroughly validated through rigorous testing in peer-reviewed literature. When evaluating MM-Fi's practical implementation, we consider both power and computational requirements. According to device specifications, LiDAR consumes 6.50 W during operation and 0.30 W in standby mode, while WiFi requires 1.21 W and 0.01 W respectively. Computationally, LiDAR demands 73.44 GFLOPS during active use compared to WiFi's 31.64 GFLOPS, with both requiring negligible resources in standby. To thoroughly assess performance across configurations, we develop and test multiple models: one utilizing all five modalities, separate models for each possible four-modality combination, and a specialized "RGB+Li+Wi" configuration.

Fig. 27 demonstrates the performance of various modality combinations in the MM-Fi dataset. The expansion of the HAR task to encompass 27 distinct classes results in an overall accuracy decrease, as illustrated in Fig. 27(a). Though both depth sensors and LiDAR are imaging modalities, removing depth sensors causes more significant accuracy degradation since they provide more comprehensive textural information of human body surfaces, enabling more accurate recognition network determinations. Similarly, while radar and WiFi are both RF devices with penetration capabilities utilizing the Doppler effect, radar elimination has a more detrimental impact on accuracy for reasons detailed in §5.6. Fig. 27(b) shows that introducing LiDAR and WiFi increases costs. Although these additions confirm the E-M^2's generalizability and marginally enhance accuracy, these improvements don't justify the additional resource consumption, validating our original modality selection.

6 Conclusion

In this study, we have introduced E-M^2, a significant step towards enhancing the efficiency of AIoT systems. E-M^2 effectively reduces both power consumption and computational overhead by adaptively activating sensors and computing modules. By employing a novel policy network to minimize modality redundancy and underutilization, and dealing with time variation and modality anomalies, E-M^2 enhances the efficiency of multimodal systems without compromising sensing performance, thereby enabling their seamless integration into edge devices for broader adoption. Extensive experiments across various sensing tasks, modalities, and environments have underscored E-M^2 's promising performance in efficient multimodal sensing. Looking forward, we plan to collaborate with industry partners to incorporate our technology into consumer electronics, driving wider acceptance in real-world applications.

Acknowledgments

The study is supported by Shenzhen Science and Technology Program (No. 20231120215201001) and National Natural Science Foundation of China (No. 62502191).

REFERENCES

[1] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal Ambiguity Learning for Multimodal Fake News Detection," in *Proc. of ACM WWW*, 2022, pp. 2897–2905.

[2] H. Xu, Z. Yang, Z. Zhou, L. Shangguan, K. Yi, and Y. Liu, "Indoor Localization via Multi-modal Sensing on Smartphones," in *Proc. of ACM UbiComp*, 2016, pp. 208–219.

[3] Y. Weng, G. Wu, T. Zheng, Y. Yang, and J. Luo, "Large Model for Small Data: Foundation Model for Cross-modal RF Human Activity Recognition," in *Proc. of the 22nd ACM SenSys*, 2024, pp. 436–449.

[4] Y. Weng, T. Zheng, Y. Yang, and J. Luo, "FM-Fi 2.0: Foundation Model for Cross-Modal Multi-Person Human Activity Recognition," *IEEE Transactions on Mobile Computing*, 2025.

[5] X. Wang, L. Kong, T. Wei, L. He, G. Chen, J. Wang, and C. Xu, "VLD: Smartphone-assisted Vertical Location Detection for Vehicles in Urban Environments," in *Proc. of the 19th ACM/IEEE IPSN*. IEEE, 2020, pp. 25–36.

[6] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019.

[7] X. Guo, J. Liu, and Y. Chen, "FitCoach: Virtual Fitness Coach Empowered by Wearable Mobile Devices," in *Proc. of IEEE INFOCOM*. IEEE, 2017, pp. 1–9.

[8] P. Zhang, M. Rostami, P. Hu, and D. Ganesan, "Enabling Practical Backscatter Communication for On-body Sensors," in *Proc. of the 2016 ACM SIGCOMM*, 2016, pp. 370–383.

[9] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart Homes that Monitor Breathing and Heart Rate," in *Proc. of the 33rd ACM CHI*, 2015, pp. 837–846.

[10] D. Vasisht, A. Jain, C.-Y. Hsu, Z. Kabelac, and D. Katabi, "Duet: Estimating User Position and Identity in Smart Homes using Intermittent and Incomplete RF-data," *Proc. of the ACM UbiComp*, vol. 2, no. 2, pp. 1–21, 2018.

[11] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing Through Fog without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather," in *Proc. of the IEEE/CVF CVPR*, 2020, pp. 11 682–11 692.

[12] M. I. Daepp, A. Cabral, V. Ranganathan, V. Iyer, S. Counts, P. Johns, A. Roseway, C. Catlett, G. Jancke, D. Gehring *et al.*, "Eclipse: An End-to-end Platform for Low-cost, Hyperlocal Environmental Sensing in Cities," in *Proc. of the 21st ACM/IEEE IPSN*. IEEE, 2022, pp. 28–40.

[13] A. Curtis, A. Pai, J. Cao, N. Moukaddam, and A. Sabharwal, "HealthSense: Software-defined Mobile-based Clinical Trials," in *Proc. of the 25th ACM MobiCom*, 2019, pp. 1–15.

[14] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, "EmotionSense: a Mobile Phones Based Adaptive Platform for Experimental Social Psychology Research," in *Proc. of the ACM UbiComp*, 2010, pp. 281–290.

[15] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Audiovisual Information Fusion in Human–computer Interfaces and Intelligent Environments: A Survey," *Proc. of the IEEE*, vol. 98, no. 10, pp. 1692–1715, 2010.

[16] S. Zhang, T. Zheng, Z. Chen, J. Hu, A. Khamis, J. Liu, and J. Luo, "OCHID-Fi: Occlusion-robust Hand Pose Estimation in 3D via RF-vision," in *Proc. of the IEEE/CVF ICCV*, 2023, pp. 15 112–15 121.

[17] Verkada. (2021) Sensor Power Requirements. Accessed: 2023-10-05. [Online]. Available: https://help.verkada.com/en/articles/5589645-camera-power-requirements

[18] K. Wang, J. Cao, Z. Zhou, and Z. Li, "SwapNet: Efficient Swapping for DNN Inference on Edge AI Devices Beyond the Memory Budget," *IEEE Transactions on Mobile Computing*, 2024.

[19] J. Sorber, A. Balasubramanian, M. D. Corner, J. R. Ennen, and C. Qualls, "Tula: Balancing Energy for Sensing and Communication in a Perpetual Mobile System," *IEEE Transactions on Mobile Computing*, vol. 12, no. 4, pp. 804–816, 2012.

[20] S. Kang, J. Lee, H. Jang, Y. Lee, S. Park, and J. Song, "A Scalable and Energy-efficient Context Monitoring Framework for Mobile Personal Sensor Networks," *IEEE Transactions on Mobile Computing*, vol. 9, no. 5, pp. 686–702, 2009.

[21] H. Pan, F. Tan, Y.-C. Chen, G. Huang, Q. Li, W. Li, G. Xue, L. Qiu, and X. Ji, "DoCam: Depth Sensing with an Optical Image Stabilization Supported RGB Camera," in *Proc. of the 28th ACM MobiCom*, 2022, pp. 405–418.

[22] Z. Xie, X. Ouyang, L. Pan, W. Lu, X. Liu, and G. Xing, "HiToF: a ToF Camera System for Capturing High-resolution Textures," in *Proc. of the 28th ACM MobiCom*, 2022, pp. 764–765.

[23] T. Zheng, Z. Chen, C. Cai, J. Luo, and X. Zhang, "V2iFi: In-vehicle Vital Sign Monitoring via Compact RF Sensing," *Proc. of the ACM UbiComp*, vol. 4, no. 2, pp. 1–27, 2020.

[24] Z. Chen, T. Zheng, and J. Luo, "Octopus: A Practical and Versatile Wideband MIMO sensing Platform," in *Proc. of the 27th ACM MobiCom*, 2021, pp. 601–614.

[25] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand, "Capturing the Human Figure through a Wall," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–13, 2015.

[26] E. K. Naeini, S. Shahhosseini, A. Kanduri, P. Liljeberg, A. M. Rahmani, and N. Dutt, "AMSER: Adaptive Multimodal Sensing for Energy Efficient and Resilient eHealth Systems," in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2022, pp. 1455–1460.

[27] S. Hor, M. El-Khamy, Y. Zhou, A. Arbabian, and S. Lim, "CM-ASAP: Cross-modality Adaptive Sensing and Perception for Efficient Hand Gesture Recognition," in *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2024, pp. 207–213.

[28] R. Panda, C.-F. R. Chen, Q. Fan, X. Sun, K. Saenko, A. Oliva, and R. Feris, "AdaMML: Adaptive Multi-modal Learning for Efficient Video Recognition," in *Proc. of the IEEE/CVF ICCV*, 2021, pp. 7576–7585.

[29] Z. Xue and R. Marculescu, "Dynamic Multimodal Fusion," in *Proc. of the IEEE/CVF CVPR*, 2023, pp. 2575–2584.

[30] O. Mees, A. Eitel, and W. Burgard, "Choosing Smartly: Adaptive Multimodal Fusion for Object Detection in Changing Environments," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2016, pp. 151–156.

[31] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, "Unravelling Robustness of Deep Learning Based Face Recognition Against Adversarial Attacks," in *Proc. of AAAI*, vol. 32, no. 1, 2018.

[32] P. J. Werbos, "Backpropagation Through Time: What It Does and How to Do It," *Proc. of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[33] A. Javaloy, M. Meghdadi, and I. Valera, "Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 9938–9964.

[34] T. Zheng, A. Li, Z. Chen, H. Wang, and J. Luo, "AutoFed: Heterogeneity-aware Federated Multimodal Learning for Robust Autonomous Driving," in *Proc. of the 29th ACM MobiCom*, 2023, pp. 1–15.

[35] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal Fusion for Multimedia Analysis: a Survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of the IEEE/CVF CVPR*, 2016, pp. 770–778.

[37] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep Reinforcement Learning that Matters," in *Proc. of AAAI*, vol. 32, no. 1, 2018.

[38] S. Hochreiter, "Long Short-term Memory," *Neural Computation MIT-Press*, 1997.

[39] E. Jang, S. Gu, and B. Poole, "Categorical Reparameterization with Gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.

[40] E. J. Gumbel, *Statistical Theory of Extreme Values and Some Practical Applications: a Series of Lectures*. US Government Printing Office, 1954, vol. 33.

[41] G. Hinton, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, 2015.

[42] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[43] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-propagating Errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[44] K. Cho, "Learning Phrase Representations using RNN Encoder-decoder for Statistical Machine Translation," *arXiv preprint arXiv:1406.1078*, 2014.

[45] J. Hadamard, "Étude sur les Propriétés des Fonctions Entières et en Darticulier d'une Fonction Considérée par Riemann," *Journal de Mathématiques Pures et Appliquées*, vol. 9, pp. 171–215, 1893.

[46] Novelda AS. (2017) Single-Chip Radar Sensors with Sub-mm Resolution. Accessed: 2024-05-13. [Online]. Available: https://www.xethru.com/
[47] Vzense. (2022) Vzense ToF Camera Evaluate Tool For Windows System. Accessed: 2024-05-13. [Online]. Available: https://github.com/Vzense/UTool
[48] NVIDIA Corporation. (2022) NVIDIA GeForce RTX 4090. Accessed: 2024-05-05. [Online]. Available: https://www.nvidia.com/en-us/geforce/graphics-cards/40-series/rtx-4090/
[49] NVIDIA Developer. (2024) Jetson TX2 NX Developer Kit with Official core module for Learning AI Programming. Accessed: 2024-05-13. [Online]. Available: https://category.yahboom.net/products/tx2-nx
[50] WASITES. (2013) WASITES PZ9002 Digital Power Meter. Accessed: 2024-05-23. [Online]. Available: https://www.hzk17.com/page96?product_id=61
[51] NVIDIA Corporation. (2023) NVIDIA Nsight Tools. Accessed: 2024-05-05. [Online]. Available: https://developer.nvidia.com/nsight-developer-tools/
[52] F. Developers. (2023) FFmpeg. Accessed: 2024-05-05. [Online]. Available: https://ffmpeg.org/
[53] Z. Zhao, C. Liu, X. Guang, and K. Li, "MLRS-RL: An Energy-efficient Multilevel Routing Strategy based on Reinforcement Learning in Multimodal UWSNs," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 708–11 723, 2023.
[54] H. Gupta, V. Navda, S. Das, and V. Chowdhary, "Efficient Gathering of Correlated Data in Sensor Networks," *ACM Transactions on Sensor Networks (TOSN)*, vol. 4, no. 1, pp. 1–31, 2008.
[55] Livox. (2025) Livox Mid-360 LiDAR. Accessed: 2024-03-15. [Online]. Available: https://www.livoxtech.com/mid-360
[56] ASUS. (2025) ASUS RT-AX86U Pro. Accessed: 2024-03-15. [Online]. Available: https://www.asus.com/fi/networking-iot-servers/wifi-routers/asus-gaming-routers/rt-ax86u-pro
[57] J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "MM-Fi: Multi-modal Non-intrusive 4d Human Dataset for Versatile Wireless Sensing," *Proc. of NeurIPS*, vol. 36, 2024.

Jinyi Cui is a Master's student at the Southern University of Science and Technology (SUSTech). He received his Bachelor's degree from Southern University of Science and Technology in 2024. His research interests include mobile computing, RF sensing, multimodal sensing, and machine learning.

Tianyue Zheng is an Assitant Professor and Ph.D. Supervisor at the Southern University of Science and Technology (SUSTech). He received his Ph.D. degree from Nanyang Technological University, Singapore, in 2023. His research interests focus on mobile computing, RF sensing, and multimodal sensing. He has published over 30 papers in top venues. He serves program committee members for several international conferences and a reviewer for multiple journals.