Towards Efficient Heterogeneous Multi-Modal Federated Learning with Hierarchical Knowledge Disentanglement

Xingchen Wang, Haoyu Wang, Feijie Wu, Tianci Liu, Qiming Cao, Lu Su*
Purdue University, West Lafayette, IN, USA
{wang2930,wang5346,wu1977,liu3351,cao393,lusu}@purdue.edu

Abstract

Multi-modal sensing systems are becoming increasingly common in real-world applications like human activity recognition (HAR). To enable knowledge sharing among individuals, Federated Learning (FL) offers a solution as a distributed machine learning paradigm that retains user data locally, thereby safeguarding privacy. However, existing heterogeneous multi-modal Federated Learning (MMFL) solutions have yet to fully utilize all the potential knowledge-sharing opportunities, as they fail to capture fundamental common knowledge that is independent of both modality and client. In this paper, we propose Federated Hierarchical Knowledge Disentanglement (FedHKD), a new sensing system for heterogeneous multi-modal federated learning. FedHKD introduces a multi-stage training paradigm based on hierarchical knowledge disentanglement at both the modality and client levels. This design enhances collaboration among modality-heterogeneous clients while maintaining low storage overhead and high adaptation flexibility to new sensing modalities. Our evaluation of two public real-world multi-modal HAR datasets and a self-collected dataset demonstrates that FedHKD outperforms state-of-the-art baselines by up to 4.85% in accuracy while saving up to 2.29× in storage. Additionally, when adapting to new sensing modalities, it reduces communication overhead by up to 4.62×.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing**; • **Computing methodologies** → **Learning paradigms**.

Keywords

Multi-modal model, Federated learning, Modality heterogeneity, Knowledge disentanglement, Parameter-efficient fine-tuning

ACM Reference Format:
Xingchen Wang, Haoyu Wang, Feijie Wu, Tianci Liu, Qiming Cao, Lu Su. 2024. Towards Efficient Heterogeneous Multi-Modal Federated Learning with Hierarchical Knowledge Disentanglement. In *ACM Conference on Embedded Networked Sensor Systems (SenSys '24), November 4–7, 2024, Hangzhou, China.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3666025.3699360

*Lu Su is the corresponding author.

SenSys '24, November 4–7, 2024, Hangzhou, China

ACM ISBN 979-8-4007-0697-4/24/11
https://doi.org/10.1145/3666025.3699360

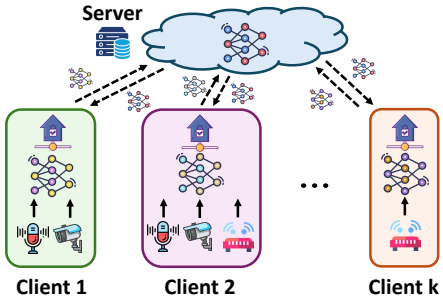


Figure 1: A typical heterogeneous multi-modal federated learning system.

1 Introduction

Multi-modal sensing systems leverage data from diverse sensors (e.g., cameras, microphones, LiDAR, WiFi, mmWave, etc.) to provide comprehensive insights into various phenomena [9, 12, 17, 20, 53]. These systems are integral to numerous applications, particularly in human activity recognition (HAR) [31, 32, 34], where sensor data are fused to generate accurate and robust models for activity prediction. By combining information from diverse sources, multi-modal sensing systems enhance the depth and breadth of analysis, facilitating a deeper understanding of complex real-world scenarios. However, a single client's data is usually limited, preventing the model from achieving optimal performance. Fortunately, Federated Learning [10, 19, 26, 30, 42–44], allows clients to share knowledge while keeping all raw data local, thereby preserving privacy. In a multi-modal federated learning (MMFL) system, our goal is to provide an optimal local model for each client, taking multi-modality data as input and predicting activity labels.

In practice, as shown in Figure 1, it is common for clients to have different sets of sensors due to sensor failures or limited accessibility to certain modalities, leading to modality heterogeneity among clients [32, 58, 60]. This poses a significant design challenge for federated learning, as it must effectively handle diverse multi-sensor clients while maintaining high performance.

Several previous attempts have been made in the field of heterogeneous MMFL [32, 47, 59, 60]. These can generally be categorized into three main groups: (1) *imputation-based FL* [60], which utilizes data imputation to handle modality heterogeneity. However, [60] necessitates the use of public data for pretraining a modality imputation model. Finding suitable public data for a specific sensor set, particularly in wireless sensing tasks involving mmWave radar, ultrasonic signals, or WiFi signals, may not always be feasible. (2) *Client-wise FL* [47] restricts the federated learning to clients with the same set of sensing modalities. However, this approach limits training to clients with identical modality sets, thus missing the opportunity to collaborate with a broader range of clients. (3) *modality-wise*

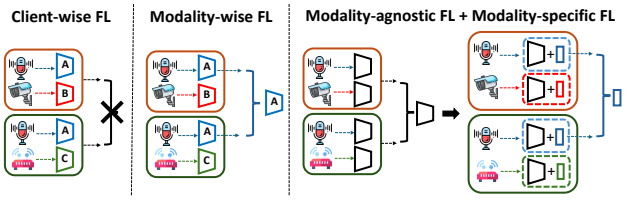


Figure 2: Typical paradigms on heterogeneous multi-modal FL systems.

FL [32, 59], which aggregates feature encoders at the modality level, allowing all feature encoders for the same sensing modality to be aggregated regardless of client-level matching. Though the restrictions are alleviated, they still impose limitations on collaboration scope. As illustrated in Figure 2, consider a typical heterogeneous MMFL system: one client possesses sensors A and B, while the other has sensors A and C. Due to their differing sensor sets, *client-wise FL* is not applicable in this scenario. Although *modality-wise FL* can collectively learn feature encoder A, it still does not fully exploit all potential for collaborative learning between these clients. In fact, there is common knowledge for the same activity that is independent of modality or client. Both modality B and modality C contain such common knowledge, which should be included in the collaboration. By incorporating this modality-agnostic knowledge, effective sharing can occur between the modality-heterogeneous clients.

To mitigate the identified limitations while simultaneously delivering personalized solutions to users, we propose a two-level knowledge disentanglement process, where we separate information within each modality into modality-agnostic and modality-specific components, while distinguishing between common and unique aspects among clients. Building upon this framework, we introduce a multi-stage training paradigm based on hierarchical knowledge disentanglement at both the modality and client levels. Initially, it learns modality-agnostic and client-independent knowledge using a shared base model. Subsequently, we fine-tune the model to acquire modality-specific and client-specific information sequentially. To preserve the knowledge acquired in the initial stage, we propose leveraging Low-rank Adaptation (LoRA) [16] to implement parameter-efficient fine-tuning (PEFT). This fine-tuning approach focuses solely on training a small number of parameters introduced by the modality-specific components while keeping the pretrained model unchanged, which efficiently learns modality-specific knowledge without compromising the modality-agnostic ones. Simultaneously, each client only needs to store the additional modality-specific components while sharing the same base model across all modalities. Consequently, the storage burden is significantly reduced for resource-constrained devices. Furthermore, this hierarchical design eliminates the need for complete retraining when a new sensing modality is introduced. Only the lightweight fine-tuning stage is necessary, while the modality-agnostic base model can be reused directly. This facilitates more effective and efficient handling of sensor changes in real-world scenarios. Existing approaches fail to explicitly distinguish between commonness and uniqueness among modalities. Consequently, they are unable to optimize storage by sharing a base model or efficiently adapt to new modalities with the shared base model.

To further enhance the performance of modality-specific fine-tuning, we propose a novel computational resource allocation strategy. This strategy uses the sensing quality of each modality as a measure of importance to allocate resources accordingly. To address the update heterogeneity caused by varied modality importance across clients, we introduce a dedicated aggregation mechanism. Additionally, we integrate a specialized correlation-based attentive fusion model to optimize local model performance for each client.

We implement FedHKD and conduct extensive experiments to evaluate its performance. Specifically, we compare the performance of FedHKD against four competitive baselines using three real-world multi-modal HAR datasets: two public datasets and one self-collected dataset leveraging a custom-built multi-modal testbed. Our results demonstrate that FedHKD significantly improves inference accuracy and reduces storage requirements compared to existing solutions. Additionally, when adopting a new modality not present during the training stage, FedHKD incurs considerably less communication overhead.

In summary, our contributions are as follows:

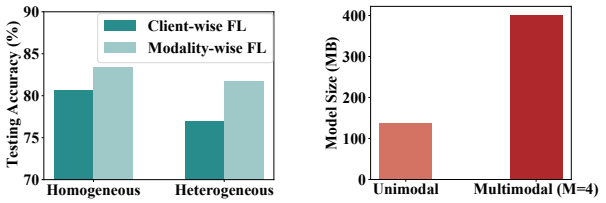
- Upon meticulous examination of modality heterogeneity in multi-modal federated learning (FL) systems, we observe that existing approaches limit collaboration among modality-heterogeneous clients to the modality level, thus leading to suboptimal knowledge sharing.
- Drawing from these insights, we introduce a novel hierarchical multi-modal federated learning framework. This framework disentangles common and unique features at both the modality and client levels, enhancing accuracy while significantly reducing storage requirements and improving adaptation to new modalities.
- To improve modality-specific fine-tuning, we propose a strategy that allocates computational resources based on sensing quality. We introduce an aggregation mechanism to handle update heterogeneity and a correlation-based attentive fusion model to optimize local model performance.
- We perform a comprehensive evaluation using two public and one self-collected multi-modal dataset. Our approach surpasses state-of-the-art baselines by up to 4.85% in accuracy, with savings of up to 2.29× in storage. Furthermore, when adapting to new sensing modalities, it reduces communication overhead by up to 4.62×.

2 Background and Motivation

We start by introducing the background of federated learning. Next, we demonstrate the necessity of enhancing performance in the presence of modality-heterogeneity, reducing model redundancy, and increasing flexibility for incorporating new modalities for an MMFL system. Additionally, we highlight the limitations of current approaches, which serve as the key motivation for our work.

2.1 Background on Federated Leaning

Federated learning [19, 30] aims to facilitate information sharing among users while preserving data privacy. In FL, a central server coordinates with numerous devices acting as clients, each equipped with a set of sensors. During each round, every device trains a local model using its own data. These clients then send their local model



(a) **Impact of modality-heterogeneity on model performance** (b) **Comparison of model sizes: unimodal vs. multimodal models**

Figure 3: Preliminary studies on modality-heterogeneity and multimodal model sizes.

updates to the central server, where the updates are aggregated to refine the global model. The updated global model is subsequently distributed back to the devices for the next round of training. This iterative process continues until the model converges. However, achieving convergence can be significantly influenced by modality heterogeneity in multi-modal federated learning [10, 26, 58], and the substantial communication overhead arising from message transmissions further hinders the convergence process in federated learning setups [24, 55].

2.2 Performance Degradation Due to Modality Heterogeneity

In a multi-modal setting, each client may have different types of sensors locally. Previous MMFL work often assumes that the multi-modal clients possess the same set of sensors [32, 47]. However, in practice, clients vary significantly in both the number and types of sensors they possess. This variability poses a substantial challenge for federated learning systems, particularly due to modality heterogeneity. When federated learning is restricted to clients with identical sensor sets, the potential for cooperation is severely limited, leading to noticeable performance degradation.

To demonstrate the impact of modality heterogeneity on model performance, we evaluate the previous methods [47] and [32] using self-collected multi-modal data (refer to Section 4.1 for dataset details). Our real-world FL testbed collects mmWave data, ultrasonic data, and depth camera data from 6 people to classify 14 human activities. We control the modality heterogeneity by assigning the modality manually. In the modality-homogeneous setup, every subject lacks ultrasonic data, while in the modality-heterogeneous one, every two subjects lack mmWave data, ultrasonic data, and depth camera data, respectively. The second setup presents higher modality-heterogeneity among the clients. As illustrated in Figure 3(a), *client-wise FL* [47] and *modality-wise FL* [32] experience approximately 4% and 2% drops in accuracy, respectively, highlighting the need to mitigate this performance degradation.

2.3 Model Redundancy across Modalities

An intuitive approach to handling multi-modal sensing data is to design separate feature encoders for each type of sensor data [32, 59, 60]. Such a method allows the model to process and learn from diverse data sources. However, it often results in redundant information duplication, as there is significant common knowledge across modalities that is learned repeatedly. Moreover, as the number of sensors increases, this strategy significantly expands the overall model size. For instance, we compared the model size between an unimodal model and a multimodal model with four sensing modalities. Utilizing separate but structurally identical transformer encoders for each modality, the multimodal model is nearly three times larger, as depicted in Figure 3(b). Meanwhile, the substantial model size impacts the federated learning process itself. It increases the communication overhead between clients and the central server, as larger models require more bandwidth and time to transmit updates. Therefore, finding efficient methods to reduce model redundancy while maintaining or even enhancing performance is crucial.

2.4 Retraining Cost for New Modality

In real-world settings, new types of sensors may become available that were not included during the initial training. Integrating such new modalities presents a significant challenge. Most existing approaches [47, 60] overlook the need to accommodate new modalities, requiring full model retraining whenever new sensor types are introduced. These approaches are both inflexible and inefficient, particularly in dynamic environments where the set of available sensors can frequently change. The need to start the training process from scratch for each modification in sensor configuration results in considerable computational overhead and time consumption. Thus, developing methods that can efficiently adapt to new sensor modalities without extensive retraining is crucial for the practical deployment of MMFL systems.

3 Methodology

3.1 Overview

We aim to develop a model that uses multi-modal data as input and predicts activity labels as output for each client. To explicitly disentangle knowledge among modalities and clients, we propose decoupling the encoder into dedicated components to separately learn common and unique features. Through this explicit knowledge disentanglement, we initially acquire modality-agnostic knowledge and subsequently fine-tune the model for modality-specific and client-specific information using low-rank adaptation (LoRA) [16]. This approach enhances collaboration among clients with diverse modalities, facilitating collaborative learning of modality-agnostic knowledge. The parameter-efficient nature of our fine-tuning method for modality-specific knowledge minimizes storage overhead. Moreover, our multi-stage training strategy simplifies the integration of new modalities, enhancing the versatility and applicability of our framework in dynamic real-world scenarios. Figure 4 illustrates an overview of the proposed framework, FedHKD, which can be mainly divided into three stages.

In the knowledge-disentangled pretraining stage, all multi-modal clients collaborate to train a modality-agnostic and client-independent model within a federated learning framework. This stage is pivotal as it establishes a foundation that can accommodate a diverse range of sensing modalities. To ensure the versatility of the encoder across different modalities, an adversarial training approach is employed. This method enables the model to discern and extract common features, regardless of the specific modality or specific client (Section 3.4).

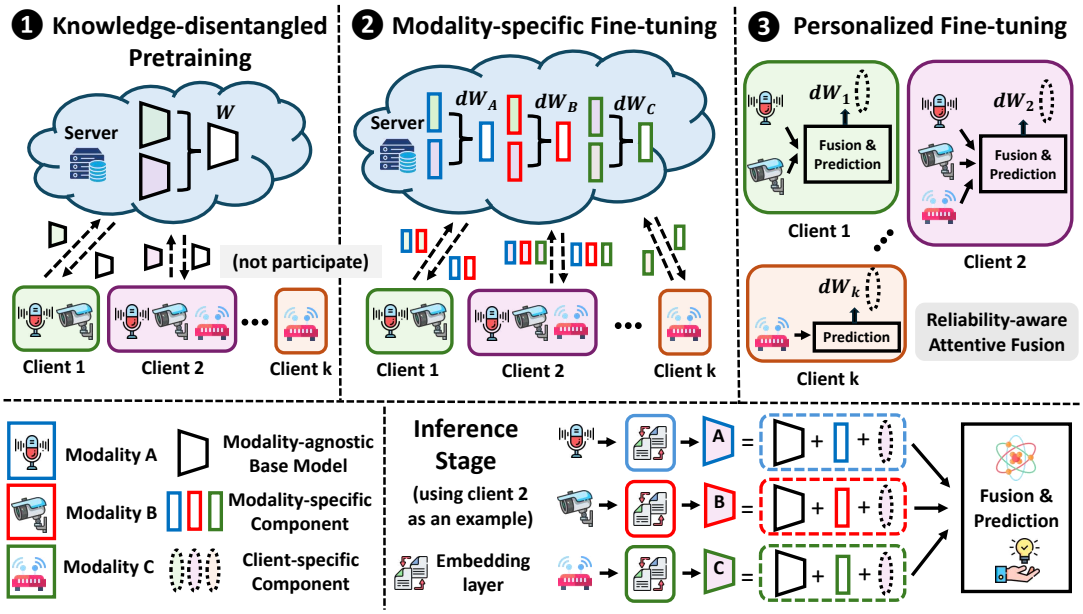


Figure 4: Overview of FedHKD framework.

The modality-specific fine-tuning stage builds upon the pre-trained model by leveraging the LoRA [16]. In this stage, we fine-tune modality-specific components, involving only a small number of parameters, which allows the model to adapt to the unique characteristics of each sensing modality. To optimize fine-tuning performance, we introduce an innovative importance-aware budget allocation technique. This mechanism efficiently allocates resources based on a dedicated importance metric, prioritizing the most significant sensing modalities given limited computational resources. Accompanying this is a heterogeneity-aware aggregation strategy achieved through client-server cooperation, ensuring seamless integration of diverse client updates. Additionally, we highlight the flexibility of our hierarchical design, demonstrating its adaptability to accommodate new modalities with minimal effort (Section 3.5).

The personalized fine-tuning stage focuses on personalization, where the local model is further refined to capture the unique characteristics of individual clients, in addition to the fundamental activity knowledge and modality-specific information. This fine-tuning process is also executed in a parameter-efficient manner, striking a balance between customization and computational efficiency. We introduce a dedicated attentive fusion mechanism designed to enhance local model performance in the multi-modal setting, enabling effective integration of information from multiple modalities (Section 3.6).

3.2 Notations

We assume that the MMFL system contains K clients covering M different data modalities. The client k maintains its own training data $D_k = \left\{\left(x_k^{(i)}, y_k^{(i)}\right)\right\}_{i=1}^{N_k}$ of size N_k, where $x_k^{(i)} = \left(x_{k,1}^{(i)}, \ldots, x_{k,M_k}^{(i)}\right)$ represents the sensing data corresponding to the i-th data sample from $M_k \leq M$ modalities, and $y_k^{(i)} \in \{1, \ldots, L\}$ is a L-way categorical label. Given that each modality has different input dimensions, we employ a lightweight linear layer for each modality to standardize these dimensions. This standardization facilitates model sharing across various modalities in the encoder component. Thus, here $x_{k,m}^{(i)} = \psi_{k,m}\left((x_{\text{raw}})_{k,m}^{(i)}\right)$ denotes the dimension-unified data, where $\psi_{k,m}$ is modality-specific linear embedding layer. In the following sections, we refer to the dimension-unified data sample as $x_{k,m}^{(i)}$ by default and a sample-label pair by (x, y) if no ambiguity arises.

3.3 Design Principle

The goal of our MMFL system is to learn a local model $\phi(\cdot)$ to predict label y given an input x, which can be an unimodal input or a multimodal input. The models are deep neural network (DNN) based that contains a feature encoder $E(\cdot)$ to extract feature $z = E(x)$ and a classification head $C(\cdot)$ acting on the extracted feature z such that $C(z)$ returns the L-length vector, with l-th entry denotes the predicted probability of label being l. Put together, we have $\phi(x) = C(E(x))$.

To explicitly decouple the commonness and uniqueness of different modalities and clients, we further decompose the feature encoder into three parts: a base encoder to capture commonness information, a modality-specific one to capture modality-unique information, and a client-specific one to model client uniqueness. Formally speaking, for client k, its extracted feature from the m-th modality sensing data of sample i is given by

$$E_{k,m}\left(x_{k,m}^{(i)}\right) = E_b\left(x_{k,m}^{(i)}\right) + E_m\left(x_{k,m}^{(i)}\right) + E_k\left(x_{k,m}^{(i)}\right).$$

Here $E_b(\cdot)$, $E_m(\cdot)$, $E_k(\cdot)$ denotes the base, modality m-specific, and client k-specific feature encoders, respectively.

To enhance the effectiveness of this design and make the proposed FedHKD lightweight, we enable the parameter-sharing and fine-tune the components E_m and E_k with the LoRA strategy [16].

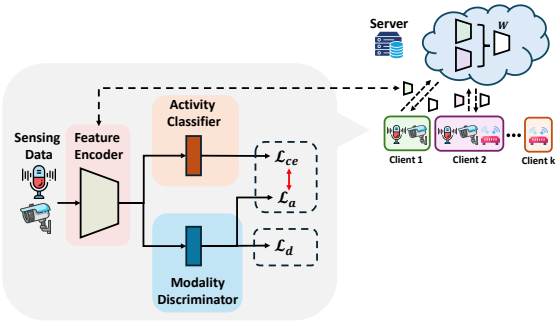


Figure 5: The Adversarial Training Architecture in Stage 1.

FedHKD is trained in a three-stage manner, as illustrated in Fig 4. We now discuss the training details.

3.4 Stage 1: Knowledge-disentangled Pretraining

This stage seeks to pre-train a good base feature encoder E_b that captures the fundamental common information of human activities, independent of modality and client, to be applicable to diverse multi-modal sensing data. This phase serves as a crucial preparatory step, laying the foundation for subsequent training stages where modality-specific and client-specific fine-tuning are introduced. The feature encoder E_b should extract *as much modality-agnostic information* as possible. To this end, we involve *all* multi-modal clients to participate in the training.

The learned features should be informative to the label, therefore, in this stage, we train the prediction model $\phi(x)$ taking E_b as the feature extractor to predict label y, i.e., $\phi(x) = C(E_b(x))$. This entails a classification task to minimize the cross-entropy (CE) loss. Specifically, for client k with its local data of size N_k, we seek to minimize

$$\mathcal{L}_{ce}(\phi) = -\frac{1}{N_k}\sum_{i=1}^{N_k}\sum_{l=1}^{L}\mathbb{1}(y^{(i)}=l)\log\phi(x^{(i)})_l$$
$$= \frac{1}{N_k}\sum_{i=1}^{N_k} CE\left(\phi(x^{(i)}), y^{(i)}\right).$$

Here $\mathbb{1}(\cdot) \in \{0,1\}$ is the indicator function. It takes 1 if the input holds valid and 0 otherwise. In addition, we denote the CE between predicted probability $\phi(x)$ and real y after one-hot encoding by $CE(\phi(x), y)$ for notation simplicity. Finally, we omit the modality and client subscripts since E_b is supposed to be unaware to them.

Loss L_{ce} suffices to train a predictive feature extractor. Notwithstanding, the desired unawareness of modality and client information cannot be ensured even if we discard them from the training process. Such failure is attributed to the potential spurious correlation in the collected data [8], and can lead to poor generalizability of E_b on tail modalities or those unseen during the training process [41]. Therefore, it is crucial to incorporate an explicit design of removing any modality- or client-specific information in this stage towards a robust pre-training.

To eliminate the modality-specific information, we resort to adversarial training [27, 40]. Specifically, as shown in Figure 5, we introduce a modality discriminator D at each client that seeks to predict the data modality from the feature extracted by E_b. Being able to fool such a discriminator indicates that the extracted features contain minimal modality information and only reflect their shared common knowledge. Motivated by this, we can train the discriminator D by minimizing:

$$\mathcal{L}_d = \frac{1}{N_k}\sum_{i=1}^{N_k} CE\left(D(E_b(x^{(i)})), one\_hot\ (m)\right),$$

where $one\_hot\ (m)$ denotes a one-hot vector with m-th entry as 1 and all other entries as 0.

Then the discriminator can predict the modality of the extracted features. To push the discriminator towards ambiguous predictions over all M modalities, we incorporate the following modality-adversarial loss during encoder training

$$\mathcal{L}_a = \frac{1}{N_k}\sum_{i=1}^{N_k} CE\left(D(E_b(x^{(i)})), \frac{\mathbf{1}_M}{M}\right), \quad (1)$$

where $\mathbf{1}_M/M$ is a vector with all entries takes $\frac{1}{M}$. In other words, after training, the discriminator will predict that any given feature as equally likely to come from all possible modalities. The local training process alternates between the modality-agnostic model (C and E_b) and the discriminator D until the convergence is achieved.

Simultaneously, client-specific information can be further minimized within a federated learning paradigm. In specific, the central server aggregates updates from all clients and combines the knowledge gained from each individual client. As a consequence, this process helps the global model cancel out the uniqueness of individual clients and is capable of achieving better generalizability thereof.

This stage ensures that encoder E_b learns to extract features devoid of modality- and client-specific information, providing a robust initialization for further training.

3.5 Stage 2: Modality-specific Fine-tuning

The base encoder E_b trained in stage 1 is capable of capturing fundamental common activity information from different modalities and clients. However, modality-specific information cannot be ignored, as it provides comprehensive insights into activity understanding, enhancing model performance.

In practice, incorporating useful modality-specific information can be challenging. First, due to the dynamic nature of the real world, the number of modalities can grow rapidly, making it prohibitive to maintain a powerful feature encoder for each modality, especially if it is of considerable size. Second, certain modalities, like LiDAR due to its high price [39], may be owned by only a few clients. This poses a significant challenge for modality-specific federated learning due to limited client participation.

To tackle the above two difficulties, we propose to obtain the modality-specific component E_m by fine-tuning the base encoder E_b, which is conducted in a parameter-efficient manner. By updating only a small number of parameters in the encoder, the updated encoder inherits much of the pre-trained knowledge. Notably, this update can be achieved with significantly less modality-specific data compared to the data required for E_b.

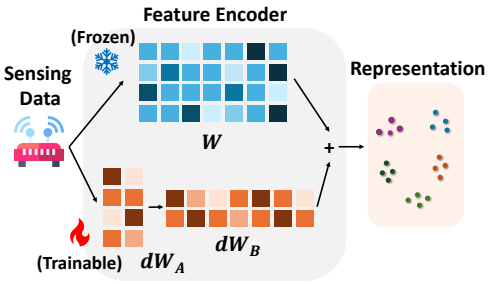


Figure 6: The working principle of modality-specific LoRA.

Our proposed parameter-efficient fine-tuning (PEFT) extends Low-rank Adaption (LoRA) [16]. In specific, LoRA models the efficient incremental updates of the pretrained model. As illustrated in Figure 6, such an efficient incremental update of pretrained model introduces two learnable low-rank matrices. Then the fine-tuned weight W' can be represented as:

$$W' = W + \Delta = W + dW_B \cdot dW_A \tag{2}$$

where $W \in \mathbb{R}^{n\times m}$ is the pretrained weights matrix, $\Delta \in \mathbb{R}^{n\times m}$ is a low-rank incremental updates matrix parameterized by $dW_A \in \mathbb{R}^{r\times m}$ and $dW_B \in \mathbb{R}^{n\times r}$ with $r \ll \min(m, n)$. During the fine-tuning, the lightweight matrices dW_A and dW_B are updated with the pre-trained weight W kept frozen.

This approach often achieves accuracy similar to that of fine-tuning the entire pre-trained model directly. By training only the small trainable components, LoRA ensures efficient computation, making it an attractive solution for fine-tuning models. More importantly, it requires only the communication of the trainable components between clients and the server, thereby reducing communication overhead in federated learning settings [45, 54].

The parameter-efficient nature of LoRA makes it suitable for learning modality-specific encoder E_m from limited data in a lightweight way. With the base encoder E_b frozen, we only need to train and communicate the modality-specific components E_m, which requires only a few parameters. However, clients usually have limited computing and storage resources, and may not afford to learn Δ with large rank r for every modality. Therefore, it is crucial to decide how to allocate the budget wisely.

Importance-Aware Budget Allocation. To tackle the allocation issue, we propose an adaptive way to allocate LoRA on each modality a proper budget (namely, the matrix rank r) based on its *importance*. Here a modality is *important* if it is of *high quality*. Inspired by previous findings that features extracted from sensing data can be decomposed into an *activity-related* part that is shared by all modalities plus some *random noises* that are specific to each modality [52], we measure the importance of a modality by how much *activity information* it carries out. If a modality is dominated by such information, we consider it important and assign more budget to learn it. In contrast, if a modality rarely agrees with others, then most of its information is likely noise, and the modality is less important. We note that the modality importance can differ across different clients, and allow each to determine its own importance, which is defined as follows.

First, we extract the modality-agnostic feature of activity sample i from client k with the pre-trained base encoder E_b by averaging

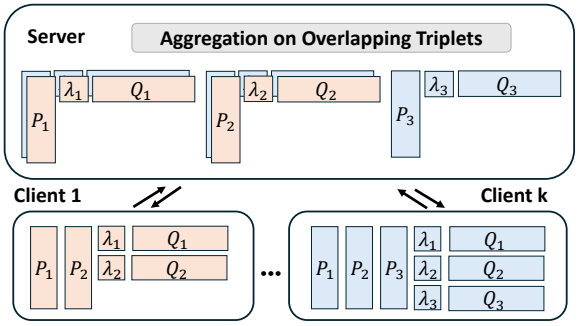


Figure 7: Illustration of heterogeneity-aware aggregation given a specific sensing modality.

features extracted from M_k modalities:

$$g_k^{(i)} = \frac{1}{M_k}\sum_{m=1}^{M_k} E_b(x_{k,m}^{(i)}).$$

In intuition, by taking an average of all modality data, $g_k^{(i)}$ smooths out the modality-specific noise and extracts the shared activity information. Built upon this measure, we check how each modality feature $E_b(x_{k,m}^{(i)})$ agrees with $g_k^{(i)}$ by their inner product

$$c_{k,m}^{(i)} = \langle E_b(x_{k,m}^{(i)}), g_k^{(i)} \rangle. \tag{3}$$

In words, $c_{k,m}^{(i)}$ quantifies the importance of modality m in activity i on client k. We next compute the importance score of each modality by normalizing $c_{k,m}^{(i)}$ over M_k modalities

$$s_{k,m}^{(i)} = \frac{c_{k,m}^{(i)}}{\sum_{m=1}^{M_k} c_{k,m}^{(i)}}.$$

Finally, we assign each modality an overall importance score on client k by averaging over all activity samples $1 \le i \le N_k$

$$s_{k,m} = \frac{1}{N_k}\sum_{i=1}^{N_k} s_{k,m}^{(i)}$$

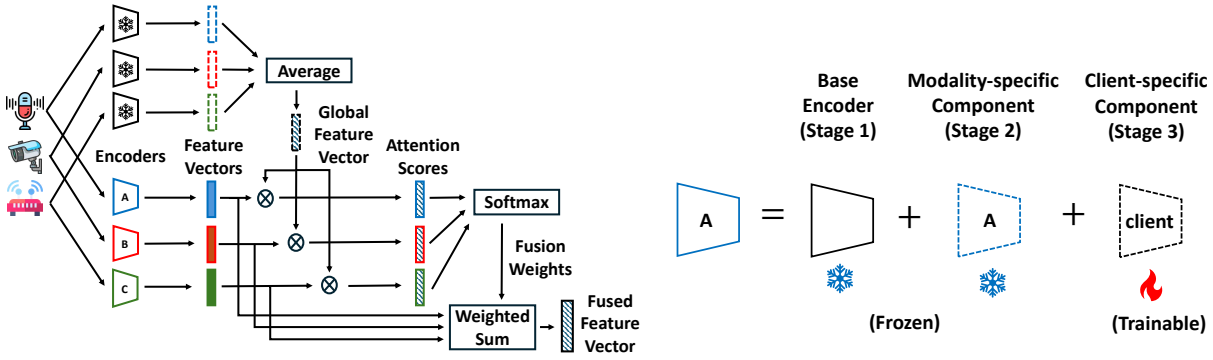
and the computation budget of client k, denoted by R, is distributed over M_k modalities. Namely, to obtain E_m on modality m, client k tunes LoRA with rank $r_m = \lfloor s_{k,m}R \rceil$, where $\lfloor\cdot\rceil$ rounds the input to the nearest interger.

Heterogeneity-aware Aggregation. According to the importance-aware allocation, a client is allowed to fine-tune each modality-aware E_m with LoRA using different ranks. Notwithstanding, these varying local rank assignments result in heterogeneous updates within each modality, making the global aggregation of modality updates from different clients challenging.

To address this issue, we propose a novel aggregation strategy. Inspired by [56], for each client, we first assign budgets (rank) to different modalities based on their local importance. Then, we parameterize the update matrices Δ using singular value decomposition (SVD). This approach enables us to aggregate local updates from different clients on a singular value-vector pair basis.

Formally, the incremental update matrix in Eqn. 2 can be expressed:

$$W' = W + \Delta = W + P \cdot \Lambda \cdot Q \tag{4}$$



(a) **Attentive Fusion** (b) **Example of Hierarchical LoRA Usage in Stage 3**

Figure 8: Illustration of proposed multi-modal fusion and hierarchical training design.

where $P \in \mathbb{R}^{n\times k}$ and $Q \in \mathbb{R}^{k\times m}$ represent the left and right singular vectors of Δ, respectively, and the diagonal matrix $\Lambda \in \mathbb{R}^{k\times k}$ contains the singular values $\{\lambda_i\}_{1\le i\le k}$. We refer to each singular value and its corresponding singular vectors as a triplet. During execution, LoRA updates are conducted by learning P, Λ, Q, respectively. To ensure the orthogonality of P and Q, we add the following regularizer:

$$R(P,Q) = \|P^T P - I\|_F^2 + \|QQ^T - I\|_F^2 \tag{5}$$

Before sending the singular vectors and singular values to the server, each client sorts the triplets based on the magnitude of the corresponding singular values. This sorting ensures a consistent order across updates. The server then aggregates these ordered triplets based on rank overlapping and distributes the aggregated values according to each client's preset ranks, as shown in Figure 7. This approach ensures efficient and effective aggregation despite the heterogeneity in local rank assignments.

Efficient Adaptation to New Modalities. In real-world applications, new modalities may be introduced in the system. Rather than retraining the entire system, our hierarchical design offers a more efficient solution. By disentangling common and unique aspects among sensing modalities, we can leverage a pre-trained modality-agnostic model as a foundation and fine-tune modality-specific delta weights for new modalities. This approach saves significant time and computational resources, ensuring the system can seamlessly accommodate new modalities.

3.6 Stage 3: Personalized Fine-tunig

After effectively fine-tuning the model for each modality, there remains significant potential for improving the system in real-world human activity recognition tasks. Typically, each client exhibits unique statures and action habits, which can degrade the performance of the global model trained through federated learning. To mitigate this issue, personalized fine-tuning emerges as an effective strategy. Similar to the modality-specific fine-tuning step, we instruct the client to freeze the shared weights E_b, modality-specific delta-weights E_m learned earlier, and focus on client-specific E_k using LoRA.

To obtain a multi-modal classifier for each client, sensor fusion has proven effective in achieving a comprehensive understanding of human activities [28, 49, 51, 52]. Among various sensor fusion methods, self-attention based fusion [52] aligns well with our goal of identifying correlations among features captured by different sensing modalities. Unlike the approach of learning a global feature vector from scratch as seen in [52], we propose a new fusion strategy tailored to our unique system design as shown in Figure 8(a). Similar to modality importance capture described in Section 3.5, we directly utilize the modality-agnostic encoder to capture the global feature representation. This eliminates the need to train the global feature vector from scratch, significantly speeding up the local fine-tuning process. However, unlike in Eqn. 3, the correlations here are calculated between the fine-tuned feature vectors and the global feature vector. The correlation for modality m of client k is as follows:

$$\begin{aligned} c_{k,m}^{(i)} &= \langle E_{k,m}(x_{k,m}^{(i)}), g_k^{(i)} \rangle \\ &= \langle E_b(x_{k,m}^{(i)}) + E_m(x_{k,m}^{(i)}) + E_k(x_{k,m}^{(i)}), g_k^{(i)} \rangle \end{aligned}$$

where the base encoder $E_b(\cdot)$ and the modality-specific component $E_m(\cdot)$ are kept frozen while the client-specific component $E_k(\cdot)$ are trainable as shown in Figure 8(b).

Next, we rescale the correlation scores using Softmax and then fuse the multi-modal features by performing a weighted sum. This parameter-efficient fine-tuning process achieves a balance between customization and computational efficiency, allowing for the effective integration of information from multiple modalities.

4 Experiment

4.1 Datasets and Models

We apply FedHKD to two representative public multi-modal HAR datasets and one self-collected dataset to show the generality of the proposed method. Table 1 summarizes the details of these datasets. Please note that these datasets are collected individually by each human subject. We adopt the data partitioning scheme introduced by FEMNIST [7] and naturally treat each subject as a client in following federated learning experiments. This approach highlights the non-IID nature of real-world applications, where user data is often skewed by user behavior.

(a) **Experiment setup** (b) **Room1** (c) **Room2** (d) **Room3**

Figure 9: Our real-world multi-modal sensor testbed for human activity recognition incorporates three sensor modalities: mmWave radar, depth camera, and smartphone for ultrasound sensing. These nodes are deployed across three distinct environments: a large conference room, a laboratory, and a small conference room.

Table 1: Statistical information of datasets (W: WiFi, M: mmWave, L: LiDAR, D: depth camera, Y: eye-tracking, G: EMG, B: body-tracking, A: acoustic)

Dataset	Sensors	# Clients	# Samples per client	# Classes
MM-Fi [50]	W+M+L+D	40	154	14
ActionSense [11]	Y+G+B	10	400	21
MMHAR	M+D+A	10	204	17

Dataset #1: MM-Fi. MM-Fi [50] is a multi-modal, non-intrusive 4D human activity dataset designed to bridge the gap between wireless sensors and high-level human perception tasks, featuring 25 categories of daily or rehabilitation actions. This dataset includes more than 320k synchronized frames across five modalities collected from 40 participants. These participants were divided into four groups evenly, each corresponding to a different environmental setting. For our evaluation, we selected 14 classes of daily activities, and four privacy-oriented modalities: WiFi, mmWave Radar, LiDAR, and depth camera. Each sample was evenly segmented into 11 units, with 8 units randomly chosen for training and the remaining 3 reserved for testing.

Dataset #2: ActionSense. ActionSense [11] is a multimodal dataset and recording framework designed to capture wearable sensing data in a kitchen setting. It features eye tracking with a first-person perspective camera, EMG sensors for forearm muscle activity, and a body-tracking system utilizing 17 inertial sensors. The dataset includes recordings of 20 different activities performed by 10 participants. Each client has at least 400 segments, with 80% allocated for training and the remaining segments left for testing.

Real-world Evaluation (MMHAR)[1] To further validate the robustness and generalizability of the proposed method, we build our own multi-modal HAR testbed by incorporating the ultrasound as an additional sensing modality and collect the data in a realistic setting. As depicted in 9(a), our experimental setup incorporates three privacy-preserving sensors: a mmWave radar [1], a depth camera [5], and a smartphone [4] (for ultrasound sensing purposes). The testbed is designed to collect synchronized multi-modal data, which is stored for further evaluation. We carefully selected these three sensors to capture human activity from diverse dimensions while preserving user privacy. Importantly, we intentionally enlarge the availability gap among the sensors. Smartphones are ubiquitous in daily life, while the other two sensors may be less common, reflecting the issue of modality missing in our daily lives. This deliberate variation in sensor availability allows us to explore and address challenges related to modality-heterogeneity problems.

[1]All the data collection was approved by IRB of the authors' institution.

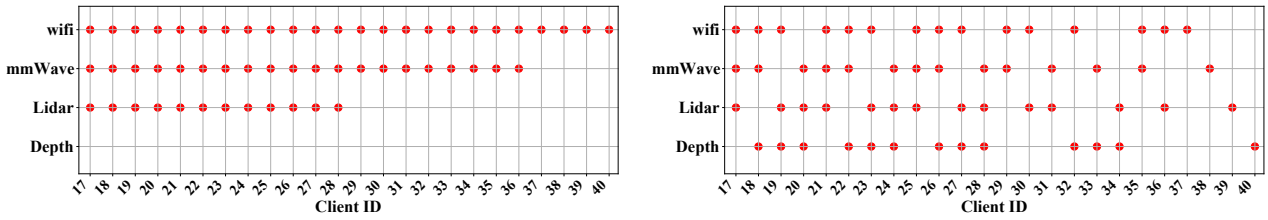
Table 2: Details of the Transformer architecture.

Hyperparameters	Values
# Layers	3
# Attention heads	2
Model dimension	768
Feed-forward network hidden dimension	768*2

To capture the main moving part of the human subject, we position the depth camera and smartphone on two tripods. The three sensing devices are placed at similar heights ranging from 1.1 meters to 1.4 meters, which may vary depending on the specific experimental environment. Ten subjects participated in the data collection, conducting activities in different rooms: four in room 1, three in room 2, and three in room 3. The multi-modal data are synchronized using the system clock and annotated using depth videos. The sampling rates of the mmWave radar, depth camera, and smartphone are 44.1 kHz, 30 Hz, and 44.1 kHz, respectively. We preprocess each sensing modality data separately. For radar data, we use a series of FFT preprocessing steps [35] to generate time-doppler heatmaps. For ultrasound signals, upon receiving the reflected pure tone signal, we first demodulate the acoustic signals using a coherent detector [38], then apply STFT on the processed data to generate DFS profiles containing velocity information. Finally, we segment all data into 3-second time windows with synchronized timestamps to obtain paired multi-modal data. This results in 12 segments per class, further divided into 9 for training and 3 for testing.

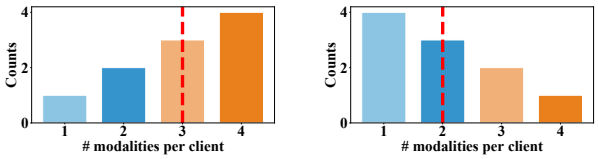
Model. FedHKD aims to enable parameter sharing across different modalities. To achieve this, we need a unified backbone neural architecture for all modalities. Given the Transformer's dominant performance across various modalities and its proven effectiveness in previous multi-modal studies [21, 29, 37], we choose it as the primary component of our model, which consists of three identical layers. The detailed architecture is outlined in Table 2.

Since the data from each modality have different dimensions, we add a lightweight embedding layer for each modality before passing the data into the encoder. These modality-specific embedding layers



(a) **Low Modality Heterogeneity** (b) **High Modality Heterogeneity**

Figure 10: Modality distribution in MM-Fi dataset for two levels of modality heterogeneity setups (client #1 to #16 with all modalities).



(a) **Mild Modality Missing** (b) **Severe Modality Missing**

Figure 11: Histograms of modality number per client in MM-Fi dataset for two levels of modality missing setups (per every ten clients). The red line denotes the average number of modalities per client.

are trained locally and not involved in federated learning, thus avoiding any additional communication overhead. Initially, these layers are trained from scratch, and in subsequent stages, they are fine-tuned in a full fine-tuning manner. As part of a multi-stage training scheme, the classifiers are retrained multiple times, as they are essential for assisting the training of the encoders. To reduce the unnecessary training workload associated with the classifiers, similar to the embedding layers, they do not need to be retrained from scratch each time. In Stage 2, each modality fine-tunes its classifier using the base classifier developed in the first Stage. In the last Stage, the modality-specific classifiers are averaged, and the resulting classifier is then slightly fine-tuned.

4.2 Experimental Setup

Baselines. To demonstrate the effectiveness of FedHKD, we compare it with the following baseline methods, each providing a distinct perspective on heterogeneous MMFL sensing systems:

- **Standalone:** In this approach, each client independently trains its model using only its local data, without any collaboration with other clients. This method serves as a basic benchmark to evaluate the performance of collaborative methods against individual learning efforts.
- **Client-wise FL [47]:** Client-wise FL restricts collaboration to clients with identical sets of sensing modalities, by focusing on homogeneous groups of clients.
- **Modality-wise FL [59]:** Modality-wise FL extends collaboration beyond homogeneous groups by aggregating all feature encoders for the same sensing modality to be aggregated regardless of client-level matching (e.g., between a uni-modal client and a multi-modal client).
- **Harmony [32]:** Building upon the principles of Modality-wise FL, Harmony incorporates additional collaborative training on the classifier to mitigate potential biases introduced during feature fusion. These additional collaborations are limited to multi-modal clients with the same set of sensors.

For a fair comparison, all baselines will utilize the same encoder backbone as ours.

Heterogeneous MMFL Setup. By default, for the MM-Fi dataset, the client distribution is proportionally set at 4:3:2:1 for clients with all modalities, and those missing one, two, and three modalities, respectively. For the other two datasets, which have only three sensing modalities, the client distribution is proportionally set at 4:3:3 for clients with all modalities, and those missing one and two modalities, respectively.

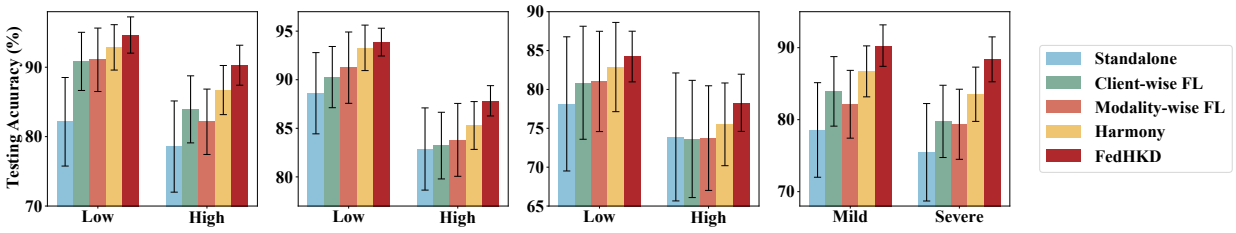
To comprehensively study the heterogeneous MMFL scenarios, we design the distribution of modalities among clients to vary at two orthogonal levels:

- **Level of Modality Heterogeneity.** In a heterogeneous MMFL system, a client may miss some sensing modalities locally, either missing the same modality or different ones. Therefore, we design two levels of modality heterogeneity. In the low modality-heterogeneous case, clients miss the same type of modalities given a specific number of missing modalities. In the high modality-heterogeneity case, clients miss different types of modalities, which are randomly assigned, as shown in Figure 10.
- **Level of Modality Missing.** Besides varying in modality heterogeneity, the severity of modality missing also impacts performance. To study this factor, we set two levels of modality missing as shown in Figure 11. Figure 11(a) represents the mild modality missing case, where the number of clients follows the ratio of 4:3:2:1 for clients with all modalities, and those missing one, two, and three modalities, respectively. This mild case results in an average of three sensing modalities per client. Conversely, we set the ratio to 1:2:3:4 for the severe modality missing case, leading to an average of two sensing modalities per client.

Unless otherwise specified, evaluations are conducted under conditions of low modality heterogeneity and mild modality missing.

4.3 System Implementation

Given that our goal is to provide an optimal local model for each client, and considering the limited size of the datasets we use, we include all clients in the federated learning training process, similar to the approach taken by our baseline [32]. We adopt FedAvg [30]



(a) **MM-Fi** (b) **ActionSense** (c) **MMHAR** (d) **MM-Fi (Different Modality Missing Level)**

Figure 12: Local model performance on different heterogeneous MMFL setups (Low / High: Low / High modality-heterogeneity; Mild / Severe: Mild / Severe modality missing).

as our aggregation scheme, as it is the most widely used method in federated learning.

During the federated learning process, each training client conducts 5 local training epochs per communication round during stages 1, and 2 local training epochs for stage 2. The neural network is implemented using PyTorch [33] and trained using the Adam optimizer [22]. The learning rates for the three training stages are set to 10^{-3}, 10^{-3}, and 10^{-5}, respectively. The LoRAs in the final two stages are trained with an average rank of 64 for each modality. Training is performed with a batch size of 16 on a server with NVIDIA A6000 GPU [3] and Intel Xeon Gold 6254 CPU [2]. The inference is conducted on the same machine. Each experiment is repeated three times using different random seeds, and the averaged results are presented in the following results sections. We report the best test accuracy of our method and baselines in 100 communication rounds.

4.4 Overall Performance

4.4.1 Impact of Modality-heterogeneity Level on Performance. First, we compare the inference accuracy of FedHKD with all the baselines in both low and high-modality-heterogeneous situations using different datasets. FedHKD consistently outperforms the baselines in both scenarios. Specifically, in the low modality-heterogeneity setting, the performance with the standard deviation across clients in Figure 12(a), 12(b), and 12(c) shows that FedHKD improves inference accuracy of the highest-performing baseline by 1.78%, 0.59%, and 1.35% on MM-Fi, ActionSense, and MMHAR, respectively. When modality heterogeneity increases, the performance gaps become more pronounced, with improvements of 3.58%, 2.54%, and 2.77% on these datasets, respectively. These results demonstrate the effectiveness of FedHKD in handling heterogeneous MMFL systems. Furthermore, the larger performance gaps in high modality-heterogeneity scenarios indicate that the commonality among modalities is underutilized in existing methods for modality-heterogeneous clients. From a dataset perspective, FedHKD enhances inference accuracy more on MM-Fi compared to the other two datasets due to the greater number of sensing modalities in MM-Fi. As the number of sensing modalities in the system increases, the likelihood of higher modality-heterogeneity also rises. With more modalities, FedHKD's advantages are more pronounced. For the following evaluations, all results are tested on the MM-Fi dataset.

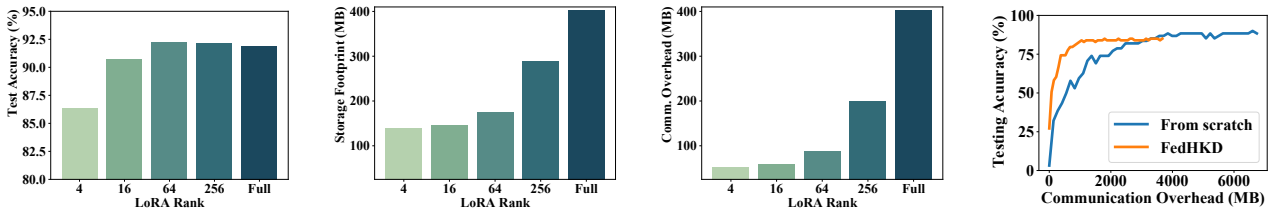
4.4.2 Impact of Modality Missing Level on Performance. We compare the inference accuracy of all methods at mild and severe modality missing levels using the MM-Fi dataset while maintaining a high modality-heterogeneity setting. As shown in the Figure 12(d), FedHKD outperforms Standalone, Client-wise FL, Modlaity-wise FL, and Harmony by 11.72%, 6.36%, 8.15%, 3.58%, respectively, in mild modality missing scenarios. When more modalities are missing, the performance gaps widen to 12.91%, 8.62%, 9.02%, and 4.85%, respectively. In cases of severe modality missing, it becomes crucial for clients with more modalities to help those with fewer modalities. Addressing this challenge requires sharing modality-specific knowledge and collaboratively learning modality-agnostic knowledge for effective transfer in modality-heterogeneous environments.

4.4.3 LoRA Hyper-parameter Study. A crucial hyper-parameter in our system design is the rank of our low-rank adaptation method. A higher rank generally improves inference accuracy but also increases storage requirements and communication overhead during the federated learning process. Therefore, there is a tradeoff between effectiveness and efficiency. To determine the optimal rank that balances these factors, we compare model performance at different ranks, evaluating inference accuracy, storage footprint, and communication overhead.

Specifically, we vary the rank with values of 4, 16, 64, and 256. For comparison, we also conduct fine-tuning across the full parameter scope. The results in Figure 13(a) demonstrate that accuracy increases with a larger rank during the fine-tuning process, but this improvement reaches a point of saturation at a rank of 64. Allocating additional resources beyond this point does not yield further accuracy gains and may even cause a slight decline. This occurs because the modality-agnostic knowledge learned by the base model during the first stage is at risk of being forgotten during fine-tuning due to overfitting [23].

In terms of storage, we reduce footprint by sharing a base encoder across modalities, which is more efficient than using separate encoders. Specifically, with ranks of 4, 16, 64, and 256, the parameter-sharing model reduces the storage footprint by factors of 2.87×, 2.74×, 2.29×, 1.39×, respectively, compared to the split encoder approach.

Another advantage of adopting LoRA in a federated learning system is its communication efficiency. For the encoder, we freeze all the pretrained parameters and only train a few newly introduced parameters. Consequently, only these lightweight components need to be aggregated and distributed during federated learning, significantly reducing the communication overhead. As shown in Figure



(a) **Local model accuracy** (b) **Storage footprint** (c) **Communication per round** (d) **Adaptation to new modality**

Figure 13: Study on LoRA and new modality adaptation.

13(c), with ranks of 4, 16, 64, and 256, the LoRA-based fine-tuning method reduces communication overhead per round by factors of 7.78×, 6.84×, 4.62×, 2.01×, respectively, compared to full-model fine-tuning.

Based on the evaluation of different ranks, we ultimately selected a rank of 64 for each modality-specific fine-tuning process. This choice strikes an optimal balance, offering significant reductions in both storage footprint and communication overhead without compromising model performance. In practical deployment, storage requirements can be directly determined based on the rank value and system architecture, allowing each client to select the highest rank that it can accommodate.

4.5 New Modality Adaptation Performance

One of the key benefits of our hierarchical design is its flexibility, which proves especially valuable when incorporating a new modality into the system. Existing works such as [32, 47, 59, 60] do not adequately address this challenge; they typically require retraining the entire network from scratch or at least completely retraining the modality-specific encoder. In contrast, our approach allows for reusing the base model trained during the first stage for new modality adaptation. For a new modality that did not participate in the initial training stage, we compare fine-tuning from the pre-trained encoder to retraining the encoder from scratch. As shown in Figure 13(d), our approach converges with significantly less communication overhead while achieving similar final testing accuracy compared to the training-from-scratch method. This comparison highlights the efficiency of our hierarchical design in adaptation, which is crucial for enhancing the scalability of federated learning systems.

4.6 Ablation Study

To understand the contribution of each module in FedHKD, we perform an ablation study to assess the individual effectiveness of each component.

4.6.1 Importance-aware Budget Allocation. Firstly, we scrutinized the performance of our dynamic budget allocation strategy on LoRA in comparison to a unified budget allocation approach. The findings, detailed in Table 3 underscore the efficacy of our importance-aware budget allocation technique, showcasing a notable 1.33% enhancement in final testing accuracy. Importantly, this improvement is achieved without incurring any additional communication overhead compared to the unified budget allocation method. Furthermore, when contrasted with full model fine-tuning, our approach not only yields larger performance gains but also significantly reduces communication overhead. These results underscore the dual benefits of our design, emphasizing both its effectiveness and efficiency in the context of federated learning.

Table 3: Ablation study results.

Method	Accuracy[1] (%) (↑)	Communication overhead[2] (MB) (↓)
full model fine-tuning	93.18	402.20
unified rank assignment	93.31	**87.11**
w/o adversarial training	93.65	-
w/o attentive fusion	92.21	-
FedHKD	**94.64**	**87.11**

[1] mean accuracy of all clients' local model performance
[2] per communication round for a single client

4.6.2 Adversarial Training. To ensure the feature extracted from the base model does not include any modality-specific information, we implement adversarial training in the first stage. To demonstrate the effectiveness of this approach, we compare it to a standard training scenario that employs a unified encoder across modalities but does not incorporate adversarial training. The results show a 0.99% improvement in final testing accuracy with adversarial training. This indicates that adversarial training effectively enhances the encoder's ability to learn common knowledge shared among modalities.

4.6.3 Attentive Fusion. To understand the performance gain of our dedicated attentive fusion design, we compare our method with a simple fusion approach where the multi-modal features are averaged. It's also trained with LoRA. The results show a 2.43% improvement in final testing accuracy, indicating the efficacy of our approach in efficiently assigning weights based on modality correlation and ultimately enhancing performance.

5 Related work

Multi-modal Federated Learning Multi-modal federated learning (MMFL) [10, 26] enables model training over distributed multi-modal data without disclosing private data. However, most existing approaches do not consider the modality heterogeneity among clients [36, 47]. [59] aggregates feature encoders at the modality level. Harmony [32] disentangles training into modality-wise and federated fusing learning stages and incorporates a balance-aware resource allocation mechanism. However, they [32, 59] ignore the importance of learning modality-agnostic knowledge between

modality-heterogeneous clients, which leads to suboptimal performance in a federated learning framework. AutoFed [60] leverages multimodal sensory data through pseudo-labeling, data imputation, and client selection mechanisms. Nevertheless, their modality imputation requires the use of a public dataset which is not always feasible in real-world applications. MultimodalHD [57] encodes multimodal sensor data into high-dimensional hypervectors and uses an attentive fusion module. Nonetheless, their primary focus is on optimizing encoder-level training efficiency, which is orthogonal to our work. Moreover, their study is limited to sensor types like accelerometers and gyroscopes, which do not adequately represent severe modality heterogeneity.

Parameter-efficient Fine-tuning Parameter-efficient fine-tuning (PEFT) [13, 46, 48] typically introduces a small number of trainable parameters into pre-trained models to adapt them to specific tasks. Prompt-based methods [25] add extra soft tokens to the initial input, which is unsuitable for human activity recognition tasks and sensitive to initialization. Adapter-based methods [14, 15] inject additional trainable modules into the original frozen backbone, introducing additional computation delay during inference. LoRA and its variants [16, 56] apply low-rank matrices to approximate weight changes during fine-tuning and can merge with pre-trained weights prior to inference, thereby not adding any extra inference burden.

6 Discussion

6.1 Overhead of Adversarial Training

Adversarial training introduces additional overhead in terms of computational resources and local training time. However, the first stage of adversarial training only needs to be performed once. This effort yields significant benefits in the long run. The resulting base model established during this initial training phase serves as a solid foundation for subsequent lightweight fine-tuning. This means that rather than starting from scratch, we can leverage the pre-trained base model, making it easier and faster to adapt to new modalities or user requirements as they arise. Thus, although the initial overhead may seem considerable, it is more than justified by the long-term efficiency gains and flexibility it brings to the overall system. Ultimately, this approach allows for a more responsive and adaptable model that can effectively meet evolving demands in real-world applications.

6.2 Varying Client Participation

Clients in federated learning may contribute inconsistently to the training process due to technical limitations or user behavior. This variation in client participation across communication rounds presents a major challenge in multi-modal federated learning, particularly when dealing with modality heterogeneity. When different clients have access to different sets of sensors or modalities, it can lead to inconsistencies in the training data, complicating the learning process and potentially degrading model performance. Fortunately, our approach explicitly disentangles modality-agnostic information that can be shared among all clients, regardless of their individual modality combinations. By identifying and extracting this common knowledge, we ensure that even when some clients are unavailable or participating inconsistently, the model can still leverage this shared information. This design not only enhances the robustness of the learning process but also effectively mitigates the adverse effects associated with fluctuating client participation. As a result, the overall system maintains high accuracy and stability, even in challenging scenarios where client availability is unpredictable.

6.3 System Scalability

Scalability is a crucial aspect of a federated learning system. Our model's design of parameter-efficient fine-tuning enhances scalability by significantly reducing both communication overhead and computational costs. In the pretraining stage, as the number of clients continues to grow, it becomes essential to incorporate client selection to further improve scalability. As highlighted in Section 6.2, our method demonstrates considerable promise in accommodating sporadic client participation schemes. By designing our approach to effectively handle varying levels of client engagement, we can maintain system performance and reliability, even as the network expands. Overall, our system ensures practicality and efficiency in larger-scale deployments, ultimately paving the way for more comprehensive multi-modal federated learning applications.

6.4 Potential Applications

While this paper primarily focuses on human activity recognition, the principles and methodologies developed through multi-modal federated learning systems have broad implications across various domains beyond HAR. In autonomous vehicles, these systems can integrate data from cameras, LiDAR, and radars to enhance object detection, route planning, and driving behavior, all while preserving data privacy across vehicles [60]. Similarly, in healthcare, federated learning can support remote monitoring and personalized medicine by combining data from wearables and diagnostic tools, improving patient outcomes without compromising sensitive data [6]. In finance, multi-modal federated learning can utilize transaction records, device information, and market trends to bolster fraud detection and credit scoring while ensuring data privacy [18]. These diverse applications demonstrate the transformative potential of multi-modal federated learning to balance efficiency and privacy across a wide range of industries.

7 Conclusion

In this paper, we propose FedHKD, a novel heterogeneous multi-modal federated learning sensing system. FedHKD disentangles common and unique features at both the modality and client levels through a multi-stage training design. Extensive experiments demonstrate that FedHKD outperforms state-of-the-art baselines in accuracy and reduces storage requirements. Additionally, when adapting to new sensing modalities, it significantly reduces communication overhead.

Acknowledgments

This work is supported by the U.S. National Science Foundation under Grant CNS-2154059.

References

[1] [n. d.]. AWR1843BOOST Evaluation board | TI.com — ti.com. https://www.ti.com/tool/AWR1843BOOST. [Accessed 01-07-2024].

[2] [n. d.]. Intel® Xeon® Gold 6254 Processor (24.75M Cache, 3.10 GHz) Product Specifications — ark.intel.com. https://ark.intel.com/content/www/us/en/ark/products/192451/intel-xeon-gold-6254-processor-24-75m-cache-3-10-ghz.html. [Accessed 01-07-2024].

[3] [n. d.]. NVIDIA RTX A6000 Powered by Ampere Architecture | NVIDIA — nvidia.com. https://www.nvidia.com/en-us/design-visualization/rtx-a6000/. [Accessed 01-07-2024].

[4] [n. d.]. Samsung Galaxy S9 for business — samsung.com. https://www.samsung.com/us/business/products/mobile/phones/galaxy-s9/b2bapp/. [Accessed 01-07-2024].

[5] [n. d.]. ZED 2 - AI Stereo Camera | Stereolabs — stereolabs.com. https://www.stereolabs.com/products/zed-2. [Accessed 01-07-2024].

[6] Qiong Cai, Hao Wang, Zhenmin Li, and Xiao Liu. 2019. A survey on multimodal data-driven smart healthcare systems: approaches and applications. *IEEE Access* 7 (2019), 133583–133599.

[7] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. 2018. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097* (2018).

[8] Cristian S Calude and Giuseppe Longo. 2017. The deluge of spurious correlations in big data. *Foundations of science* 22 (2017), 595–612.

[9] Balasubramaniyan Chandrasekaran, Shruti Gangadhar, and James M Conrad. 2017. A survey of multisensor fusion techniques, architectures and methodologies. In *SoutheastCon 2017*. IEEE, 1–8.

[10] Liwei Che, Jiaqi Wang, Yao Zhou, and Fenglong Ma. 2023. Multimodal federated learning: A survey. *Sensors* 23, 15 (2023), 6986.

[11] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. 2022. ActionSense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. *Advances in Neural Information Processing Systems* 35 (2022), 13800–13813.

[12] Raffaele Gravina, Parastoo Alinia, Hassan Ghasemzadeh, and Giancarlo Fortino. 2017. Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. *Information Fusion* 35 (2017), 68–80.

[13] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608* (2024).

[14] Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *arXiv preprint arXiv:2106.03164* (2021).

[15] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*. PMLR, 2790–2799.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[17] K Huang, B Shi, X Li, X Li, S Huang, and Y Li. [n. d.]. Multi-modal sensor fusion for auto driving perception: A survey. arXiv 2022. *arXiv preprint arXiv:2202.02703* ([n. d.]).

[18] Wei Huang, Dexian Wang, Xiaocao Ouyang, Jihong Wan, Jia Liu, and Tianrui Li. 2024. Multimodal federated learning: Concept, methods, applications and future directions. *Information Fusion* 112 (2024), 102576.

[19] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. 2021. Advances and open problems in federated learning. *Foundations and trends® in machine learning* 14, 1–2 (2021), 1–210.

[20] Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. 2013. Multisensor data fusion: A review of the state-of-the-art. *Information fusion* 14, 1 (2013), 28–44.

[21] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*. PMLR, 5583–5594.

[22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[23] Max Kuhn, Kjell Johnson, Max Kuhn, and Kjell Johnson. 2013. Over-fitting and model tuning. *Applied predictive modeling* (2013), 61–92.

[24] Ang Li, Jingwei Sun, Xiao Zeng, Mi Zhang, Hai Li, and Yiran Chen. 2021. Fedmask: Joint computation and communication-efficient personalized federated learning via heterogeneous masking. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 42–55.

[25] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).

[26] Yi-Ming Lin, Yuan Gao, Mao-Guo Gong, Si-Jia Zhang, Yuan-Qiao Zhang, and Zhi-Yuan Li. 2023. Federated learning on multimodal data: A comprehensive survey. *Machine Intelligence Research* 20, 4 (2023), 539–553.

[27] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742* (2017).

[28] Shengzhong Liu, Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. 2020. Giobalfusion: A global attentional deep learning framework for multisensor information fusion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–27.

[29] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18177–18186.

[30] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[31] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 324–337.

[32] Xiaomin Ouyang, Zhiyuan Xie, Heming Fu, Sitong Cheng, Li Pan, Neiwen Ling, Guoliang Xing, Jiayu Zhou, and Jianwei Huang. 2023. Harmony: Heterogeneous multi-modal federated learning through disentangled model training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. 530–543.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[34] Sen Qiu, Hongkai Zhao, Nan Jiang, Zhelong Wang, Long Liu, Yi An, Hongyu Zhao, Xin Miao, Ruichen Liu, and Giancarlo Fortino. 2022. Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. *Information Fusion* 80 (2022), 241–265.

[35] Sandeep Rao. 2017. Introduction to mmWave sensing: FMCW radars. *Texas Instruments (TI) mmWave Training Series* (2017), 1–11.

[36] Batool Salehi, Jerry Gu, Debashri Roy, and Kaushik Chowdhury. 2022. Flash: Federated learning for automated selection of high-band mmwave sectors. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 1719–1728.

[37] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.

[38] David Tse and Pramod Viswanath. 2005. *Fundamentals of wireless communication*. Cambridge university press.

[39] Dinh Van Nam and Kim Gon-Woo. 2021. Solid-state LiDAR based-SLAM: A concise review and application. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 302–305.

[40] Haoyu Wang, Yaqing Wang, Feijie Wu, Hongfei Xue, and Jing Gao. 2023. Macular: A Multi-Task Adversarial Framework for Cross-Lingual Natural Language Understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5061–5070.

[41] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. 2017. Learning to model the tail. *Advances in neural information processing systems* 30 (2017).

[42] Feijie Wu, Song Guo, Haozhao Wang, Haobo Zhang, Zhihao Qu, Jie Zhang, and Ziming Liu. 2023. From deterioration to acceleration: A calibration approach to rehabilitating step asynchronism in federated optimization. *IEEE Transactions on Parallel and Distributed Systems* 34, 5 (2023), 1548–1559.

[43] Feijie Wu, Zitao Li, Yaliang Li, Bolin Ding, and Jing Gao. 2024. FedBiOT: LLM Local Fine-tuning in Federated Learning without Full Model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3345–3355.

[44] Feijie Wu, Xingchen Wang, Yaqing Wang, Tianci Liu, Lu Su, and Jing Gao. 2024. FIARSE: Model-Heterogeneous Federated Learning via Importance-Aware Submodel Extraction. *arXiv preprint arXiv:2407.19389* (2024).

[45] Xinghao Wu, Xuefeng Liu, Jianwei Niu, Haolin Wang, Shaojie Tang, and Guogang Zhu. 2024. FedLoRA: When Personalized Federated Learning Meets Low-Rank Adaptation. (2024).

[46] Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, and Yuntao Du. 2024. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint arXiv:2402.02242* (2024).

[47] Baochen Xiong, Xiaoshan Yang, Fan Qi, and Changsheng Xu. 2022. A unified framework for multi-modal federated learning. *Neurocomputing* 480 (2022), 110–118.

[48] Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint arXiv:2312.12148* (2023).

[49] Hongfei Xue, Wenjun Jiang, Chenglin Miao, Ye Yuan, Fenglong Ma, Xin Ma, Yijiang Wang, Shuochao Yao, Wenyao Xu, Aidong Zhang, et al. 2019. DeepFusion: A deep learning framework for the fusion of heterogeneous sensory data. In

Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing. 151–160.

[50] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. 2024. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. *Advances in Neural Information Processing Systems* 36 (2024).

[51] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th international conference on world wide web*. 351–360.

[52] Shuochao Yao, Yiran Zhao, Huajie Shao, Dongxin Liu, Shengzhong Liu, Yifan Hao, Ailing Piao, Shaohan Hu, Su Lu, and Tarek F Abdelzaher. 2019. Sadeepsense: Self-attention deep learning framework for heterogeneous on-device sensors in internet of things applications. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 1243–1251.

[53] De Jong Yeong, Gustavo Velasco-Hernandez, John Barry, and Joseph Walsh. 2021. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors* 21, 6 (2021), 2140.

[54] Liping Yi, Han Yu, Gang Wang, and Xiaoguang Liu. 2023. Fedlora: Model-heterogeneous personalized federated learning with lora tuning. *arXiv preprint arXiv:2310.13283* (2023).

[55] Mi Zhang, Faen Zhang, Nicholas D Lane, Yuanchao Shu, Xiao Zeng, Biyi Fang, Shen Yan, and Hui Xu. 2020. Deep learning in the era of edge computing: Challenges and opportunities. *Fog Computing: Theory and Practice* (2020), 67–78.

[56] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*. Openreview.

[57] Quanling Zhao, Xiaofan Yu, Shengfan Hu, and Tajana Rosing. 2024. MultimodalHD: Federated Learning Over Heterogeneous Sensor Modalities using Hyperdimensional Computing. In *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 1–6.

[58] Quanling Zhao, Xiaofan Yu, and Tajana Rosing. 2023. Attentive Multimodal Learning on Sensor Data using Hyperdimensional Computing. In *Proceedings of the 22nd International Conference on Information Processing in Sensor Networks*. 312–313.

[59] Yuchen Zhao, Payam Barnaghi, and Hamed Haddadi. 2022. Multimodal federated learning on iot data. In *2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 43–54.

[60] Tianyue Zheng, Ang Li, Zhe Chen, Hongbo Wang, and Jun Luo. 2023. Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.