Harmony: Heterogeneous Multi-Modal Federated Learning through Disentangled Model Training

Xiaomin Ouyang[1], Zhiyuan Xie[1], Heming Fu[1], Li Pan[1], Sitong Chen[1], Neiwen Ling[1], Guoliang Xing[1,*], Jiayu Zhou[2], Jianwei Huang[3,4,*]
[1]The Chinese University of Hong Kong, [2]Michigan State University, [3]The Chinese University of Hong Kong, Shenzhen, [4]Shenzhen Institute of Artificial Intelligence and Robotics for Society

ABSTRACT

Multi-modal sensing systems are increasingly prevalent in real-world applications such as health monitoring and autonomous driving. Most multi-modal learning approaches need to access users' raw data, which poses significant concerns to users' privacy. Federated learning (FL) provides a privacy-aware distributed learning framework. However, current FL approaches have not addressed the unique challenges of heterogeneous multi-modal FL systems, such as modality heterogeneity and significantly longer training delay. In this paper, we propose Harmony, a new system for heterogeneous multi-modal federated learning. Harmony disentangles the multi-modal network training in a novel two-stage framework, namely modality-wise federated learning and federated fusion learning. By integrating a novel balance-aware resource allocation mechanism in modality-wise FL and exploiting modality biases in federated fusion learning, Harmony improves the model accuracy under non-i.i.d. data distributions and speeds up system convergence. We implemented Harmony on a real-world multi-modal sensor testbed deployed in the homes of 16 elderly subjects for Alzheimer's Disease monitoring. Our evaluation on the testbed and three large-scale public datasets of different applications show that, Harmony outperforms by up to 46.35% accuracy over state-of-the-art baselines and saves up to 30% training delay.

CCS CONCEPTS

• **Human-centered computing** → **Mobile computing**; • **Computing methodologies** → **Learning paradigms**.

KEYWORDS

Multi-modal federated learning systems, Modality heterogeneity, Balance-aware resource allocation

ACM Reference Format:
Xiaomin Ouyang[1], Zhiyuan Xie[1], Heming Fu[1], Li Pan[1], Sitong Chen[1], Neiwen Ling[1], and Guoliang Xing[1,*], Jiayu Zhou[2], Jianwei Huang[3,4,*]. 2023. Harmony: Heterogeneous Multi-Modal Federated Learning through Disentangled Model Training. In *ACM International Conference on Mobile Systems, Applications, and Services (MobiSys '23), June 18–22, 2023, Helsinki, Finland.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3581791.3596844

*Guoliang Xing and Jianwei Huang are the corresponding authors.

MobiSys '23, June 18–22, 2023, Helsinki, Finland

ACM ISBN 979-8-4007-0110-8/23/06…$15.00
https://doi.org/10.1145/3581791.3596844

1 INTRODUCTION

Multi-modal sensing systems are increasingly deployed in real-world applications, such as health monitoring [34, 65], autonomous driving [24, 49] and human-computer interaction [32]. In these scenarios, the tasks are usually too complex and dynamic to accomplish based on only a single sensor modality. Incorporating information from multiple complementary data modalities provides improved model performance [46]. However, most existing studies on multi-modal learning focus on centralized processing of user's raw data [32, 65], which imposes significant privacy concerns.

As a key enabling distributed machine learning paradigm, federated learning (FL) [25, 38] has received significant attention recently, where only locally updated models (rather than raw data) are uploaded to the server. Most of the existing FL approaches focus on training *unimodal* models with only data input from a single sensor modality, such as image [28], audio [27], or inertial sensory data [53], which cannot be directly applied to heterogeneous multi-modal sensing systems. In particular, there is usually *modality heterogeneity* in multi-modal FL systems, where the sensor modalities available on different nodes vary significantly. For example, even autonomous driving systems from the same vendors may ship with different types/numbers of sensors, due to the diverse vehicle models and configurations [1]. The sensors may also fail dynamically, resulting in changing sensor modalities at runtime. As a result, model aggregation in multi-modal FL systems will be more challenging since the nodes with different modalities have significantly diverse model architectures. Moreover, such modality heterogeneity makes the model divergence between nodes much more severe than in unimodal FL systems, affecting both the accuracy and convergence of federated learning. Although several multi-modal FL approaches [47, 60, 67] allow model training over distributed multi-modal data on the nodes, most of them do not consider the coupled modality and distribution heterogeneity among the data of different nodes. Moreover, to the best of our knowledge, there is no prior work on reducing the significant convergence delay in multi-modal FL systems.

In this paper, we propose *Harmony*, a new system for heterogeneous multi-modal federated learning. Harmony adopts a modality-agnostic approach that harnesses the modality heterogeneity in multi-modal FL to achieve both high model accuracy and low training latency. The design of Harmony is motivated by the key observation that, directly aggregating unimodal encoders trained by multi-modal and single-model learning results in model performance degradation in multi-modal FL.

The key idea of Harmony is to disentangle the training of multi-modal networks in a novel two-stage framework, namely *modality-wise federated learning* and *federated fusion learning*. In modality-wise federated learning, the multi-modal nodes train multiple single-modal networks rather than a single multi-modal network. As a result, multiple *unimodal FL subsystems* run in parallel to consistently learn unimodal information. This approach naturally reduces the model divergence within a specific FL subsystem and can leverage the data of partial modalities caused by dynamic sensor failures on multi-modal nodes. Moreover, to reduce the system training latency, we propose a *dynamic resource allocation* mechanism, where the multi-modal nodes dynamically allocate resources to different single-modal training tasks to balance the delay of different unimodal FL systems. After modality-wise FL, the multi-modal nodes collaboratively learn the classifier layers through federated fusion learning. A key challenge is that the data of multi-modal nodes usually have non-iid distributions. To address this challenge, we design a novel federated fusion mechanism by *exploiting the modality biases* of different multi-modal nodes, where the server clusters the nodes according to their modality biases for model aggregation. Based on the pre-trained feature encoders, federated fusion learning converges fast and incurs only a small system overhead.

We deployed Harmony on a real-world multi-modal sensor testbed for four continuous weeks, which consists of 16 nodes installed in the homes of elderly subjects for Alzheimer's Disease monitoring. We show that Harmony can efficiently leverage three types of sensors (depth cameras, mmWave radars, and microphones) to accurately detect about a dozen of daily behaviors, despite the substantial runtime dynamics such as sensor failures. We also evaluate the performance of Harmony on three public multi-modal datasets from different applications that consist of samples of six different sensor modalities and incorporate up to 210 nodes. Our extensive evaluation shows that, Harmony significantly outperforms several existing machine learning paradigms in model accuracy and incurs less training latency under dynamic network conditions[1].

In summary, we make the following key contributions:

- We conduct an in-depth analysis and extensive evaluations of modality heterogeneity in multi-modal federated learning (FL) systems, which shows the negative impact of model aggregation between single-modal and multi-modal nodes that leads to model performance degradation in multi-modal FL.
- Building on our key findings, we propose Harmony, the first two-stage multi-modal FL framework that disentangles the training of multi-modal networks to harness the *modality heterogeneity* in multi-modal FL systems, improving both the model accuracy and convergence speed.
- To further enhance the performance of our new multi-modal FL framework, we design a new resource allocation strategy that addresses *imbalanced training delays* among different nodes and modalities, and introduce a novel federated fusion mechanism that improves model accuracy with *non-i.i.d. data distributions*.
- We implemented Harmony on a real-world multi-modal sensor testbed for Alzheimer's Disease monitoring, with nodes deployed in the homes of 16 elderly subjects for four weeks. Our experiments on the testbed and three public datasets show that, Harmony outperforms state-of-the-art baselines by up to 46.35% accuracy and saves up to 30% system training delay.

[1]The code is available at https://github.com/xmouyang/Harmony.

2 RELATED WORK

Multi-modal Learning. Multi-modal sensing systems have become prevalent in real-world applications. For example, Wavoice [31] fuses mmWave and audio signals for noise-resistant speech recognition. Liu et al. [32] integrate RFID and depth cameras for recognizing human gestures. Most work in this space is focused on the centralized approach that must gather the user's data at the central server, which imposes significant privacy concerns. Recently, Cosmo [42] proposes a cloud-edge multi-modal fusion framework for activity recognition, where the models trained on the cloud can be improved through on-device learning on the local data. However, such an approach is not designed for collaborative model training among different end users.

Unimodal Federated Learning. Federated Learning (FL) [25, 38] is a distributed machine learning paradigm that enables collaborative model training while keeping the data residing on devices. Many FL studies are proposed to address the non-i.i.d data distributions of nodes, based on the regularized term [30], post-training [63], or multi-task learning [43, 51]. There are also some studies focusing on tackling the system heterogeneity in FL, such as through client selection [29] and active sample selection [50]. However, most of the existing FL approaches are focused on training *unimodal* models with only single-modality data input, such as image [28], audio [27], or inertial sensory data[43, 53]. These approaches cannot be directly applied in multi-modal FL since the nodes with different modalities have significantly diverse model architectures. In contrast, Harmony can be applied to FL systems with heterogeneous data modalities on the nodes.

Multi-modal Federated Learning. Multi-modal federated learning allows model training over distributed multi-modal data on the nodes. However, the existing multi-modal FL approaches do not consider the coupled modality and distribution heterogeneity among the data of different nodes. For example, most of the current approaches perform multi-modal FedAvg [47, 60, 67] that directly averages the model weights of unimodal feature encoders from different nodes. However, such an approach will bring a significant performance drop when the nodes have heterogeneous modalities or data distributions. Chen *et al.* [17] focus on optimizing the aggregation weights among modalities and nodes in multi-modal FL. However, their approach FedHGB needs a validation dataset on each node and can only work for result-level multi-modal fusion. Harmony does not need any validation datasets and is applicable to feature-level fusion scheme, which is more general and can exploit the correlations and interactions between features of different modalities [11]. FedMSplit [15] employs a graph structure to capture the correlations among multi-modal networks. However, their FL framework does not consider the negative impact of aggregating feature encoders from different types of nodes. Moreover, to the best of our knowledge, we are the first to tackle the imbalanced training delays among different nodes and modalities in multi-modal FL systems, which reduces the overall system training latency.

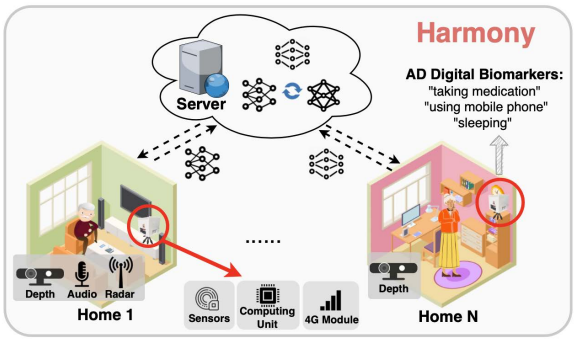


Figure 1: A typical application scenario of multi-modal federated learning systems: Alzheimer's Disease monitoring.

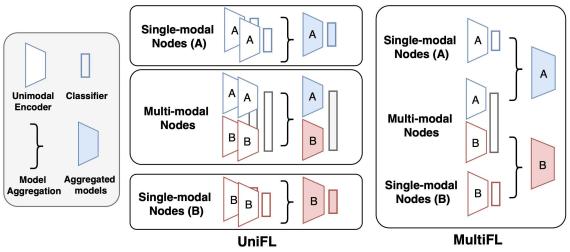


Figure 2: Two typical FL paradigms on heterogeneous multi-modal FL systems. In UniFL, only nodes with the same data modalities train models with FL. In MultiFL, single-modal and multi-modal nodes collaboratively learn models by aggregating feature encoders.

3 APPLICATIONS AND CHALLENGES

Harmony is designed for a wide class of applications where multiple heterogeneous sensors are deployed on distributed devices for running complex sensing tasks in a continuous and longitudinal manner. Representative applications include autonomous driving [49], fitness tracking [12, 13] and crowd monitoring [56].

We first briefly give two examples of the typical application scenarios of Harmony. To provide highly robust perception performance, current off-the-shelf autonomous driving cars such as Waymo Driver [6] and Baidu Apollo [3] rely on real-time fusion of multiple sensors, including cameras, lidars, radars, ultrasonic sensors, and GPS. Such sensor heterogeneity is also common in emerging smart health applications. For instance, in Alzheimer's patient monitoring scenarios [16, 22], multiple modalities of sensors (e.g., cameras, microphones, and motion sensors) are required to capture multidimensional behavior biomarkers [26], such as social interactions and physical inactivity. In the two examples, the labeled data from each node is usually limited and privacy-sensitive. As a result, the trained multi-modal models may largely suffer overfitting issues. A promising solution to address this challenge is federated learning among different nodes, which can improve the model performance in ever-changing environments, such as different road/weather conditions in autonomous driving and different users' daily routines and home layouts in Alzheimer's Disease monitoring. Next, we discuss the challenges and common practices of FL in the context of in *Alzheimer's patient monitoring*.

Figure 1 shows a multi-modal FL system for Alzheimer's patient monitoring [16, 22] that consists of nodes deployed in elderly subjects' homes. Each node is equipped with multiple modalities of sensors, such as depth cameras, mmWave radars, and microphones, to continuously track the elder's daily behaviors. For example, the nodes can detect the activities of daily living, behavioral and psychological symptoms of dementia [14], and social interactions, whose duration and frequency can be used as digital biomarkers for early AD diagnosis and intervention [2, 26]. Then, the nodes in different subjects' homes will transmit model weights to the central server to collaboratively learn the multi-modal networks while preserving local data privacy. In real-world settings, such a system would encounter many dynamics, including changing data modality and resource availability on nodes. There are already distributed sensor systems deployed in natural home environments for AD monitoring [7, 8, 37]. Previous FL studies have demonstrated that distributed learning is accurate for privacy-preserving activity recognition [43, 53], while they do not address real-world challenges in Harmony such as modality heterogeneity. In the following, we discuss these challenges in detail to motivate the design of Harmony.

Modality and data heterogeneity. In Alzheimer's patient monitoring, the data modality is usually highly heterogeneous among nodes in different subjects' homes. Such *modality heterogeneity* mainly comes from three reasons. First, due to hardware or budget constraints, some nodes may not be equipped with multi-modal sensors by design. Second, the deployment constraints such as environmental layout or privacy concerns can also make the installed sensors vary among different homes. For example, some families may not be willing to have depth cameras installed in the bedroom. Finally, the sensors may fail dynamically due to various reasons such as power surges. Moreover, different subjects usually exhibit diverse behavior patterns, resulting in *non-i.i.d. data distributions*. The coupled *modality and data heterogeneity* will bring significant model divergence among different nodes, hence posing a major challenge in designing efficient model aggregation approaches in multi-modal FL systems.

Significant training latency and overhead. The distributed nodes (e.g., mobile or edge devices) usually have very limited computing resources. Compared with conventional single-modal learning, training multi-modal networks with a larger model size on devices will incur substantially higher latency. Moreover, the wireless Internet connectivity of nodes will likely have limited and varying bandwidths, resulting in dynamic communication delays in federated learning.

4 A MOTIVATION STUDY

In this section, we evaluate the performance of current FL frameworks on nodes with heterogeneous modalities to motivate the design of Harmony.

Figure 2 illustrates the basic ideas of two typical FL frameworks in heterogeneous multi-modal FL systems. In Uni-modal Federated Learning (UniFL) [60], only nodes with the same data modalities (either single-modal or multi-modal) collaboratively learn a model. Therefore, each UniFL system only contains a subset of nodes. In Multi-modal Federated Learning (MultiFL) [17, 67], all nodes (single-modal and multi-modal) participate in the same FL system. The server averages the unimodal feature encoders from single-modal

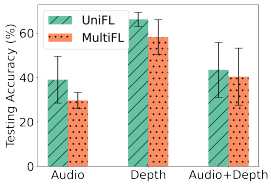


Figure 3: Accuracy of different types of nodes in UniFL and MultiFL.

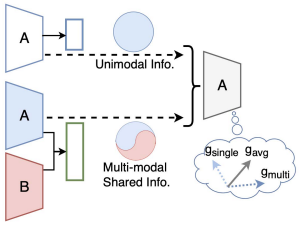


Figure 4: Information inconsistency in MultiFL.

and multi-modal nodes. Compared with UniFL, MultiFL allows model aggregation among different types of nodes, expanding the participants' scale.

Specifically, we evaluate the accuracy of UniFL and MultiFL using 16 elderly subjects' data collected by our real-world FL testbed (see Section 7.1). The task is to classify 11 behaviors related to Alzheimer's Disease using the audio and depth data. We load ten nodes with data from both modalities, three with audio data and three with depth data only. Each experiment is repeated five times. Figure 3 shows the model accuracy of nodes with different data modalities trained by UniFL and MultiFL. First, both FL paradigms have a very low model accuracy on real-world heterogeneous multi-modal data. For example, only 39% mean accuracy is achieved by the audio model trained in UniFL. Moreover, although MultiFL enables model aggregation among different types of nodes, the aggregated models of both single-modal and multi-modal nodes suffer accuracy reduction, compared with UniFL. While we only use one dataset in this section to provide an intuitive illustration of our key findings, we have obtained similar results on three other publicly available datasets, as detailed in Section 8.

To find the reasons for the accuracy drop in MultiFL, we calculate the cosine similarity of the unimodal encoder's weights from the single-modal and multi-modal networks trained on the same data. Here a smaller cosine similarity means less common information between the two models. The cosine similarity of encoders trained by single-modal and multi-modal data is 0.621 and 0.644 for audio and depth, respectively, which is relatively small. This means that there is *information inconsistency* among their unimodal encoders. As shown in Figure 4, the multi-modal network is trained to capture shared information among different modalities, which will lose some useful unimodal information. Moreover, the multi-modal network has a larger size of model parameters than a single-modal network. As a result, training the multi-modal network with limited data is more likely to encounter issues like the curse of dimensionality [39] or model overfitting [55], which can result in bad model performance. Therefore, directly averaging the encoders from multi-modal and single-modal nodes in MultiFL will have negative impacts on both sides and thus result in model performance reduction. For instance, on "Audio-only" nodes, the model accuracy after being aggregated with multi-modal nodes (i.e., 29.68% in MultiFL) is lower than that in UniFL (39.06%).

Moreover, in the presence of non-i.i.d. data distributions, the models of different multi-modal nodes may show substantial bias toward different modalities [21, 57]. For example, in Alzheimer's Disease monitoring, the subjects with dementia usually have less mobility and tend to have a sedentary lifestyle. As a result, the prediction accuracy of their multi-modal networks may rely on the encoder of depth images that captures the subjects' postures. However, cognitively normal subjects will more likely move around. Therefore, their model accuracy may depend on the encoder of radar data that captures human movements. As a result, directly averaging the models of different multi-modal nodes in both UniFL and MultiFL will yield poor accuracy.

This case study suggests two main insights. First, existing unimodal and multi-modal FL paradigms have unsatisfactory model performance on real-world heterogeneous multi-modal data. Second, compared with UniFL, the current multi-modal FL framework (e.g., MultiFL) will suffer substantial model accuracy drops on both single-modal and multi-modal nodes, due to the information inconsistency of their unimodal encoders.

5 SYSTEM OVERVIEW

We now introduce Harmony, a new system for multi-modal federated learning with heterogeneous modalities among nodes. We first introduce the problem formulation and then describe the system architecture.

5.1 Problem Formulation

Suppose there exist N nodes in the multi-modal FL system. The nodes have up to M ($M \geq 2$) different data modalities on their local data. For an arbitrary node $c_k (1 \leq k \leq N)$, its local training data set is denoted as $D_k : \{s : (\mathbf{X}, y)\}$, where $\mathbf{X} = \{\mathbf{x}_i | \forall i \in \mathcal{M}_k\}$ contains $M_k = |\mathcal{M}_k|$ $(1 \leq M_k \leq M)$ modalities. Here $\mathcal{M}_k \subseteq \{1, 2, ..., M\}$ is the valid data modalities on the node c_k. The goal of multi-modal federated learning is to learn a series of models $\{\Phi_k(\mathbf{x}_i | \forall i \in \mathcal{M}_k) | 1 \leq k \leq N\}$ for nodes with different data modalities $\mathcal{M}_k$, and minimize the training latency of the whole FL system.

Single-modal networks. A single-modal node $c_k(s)$ $(1 \leq k \leq N_s)$ trains a single-modal network $\Phi_k(s)$ based on its local modality $j \in \{1, 2, .., M\}$. The single-modal network is composed of the unimodal feature encoder $f_{enc_j}(\cdot)$ and the classifier $g_j(\cdot)$. We have:

$$\Phi_k(s) = f_{enc_j}(\cdot) \cup g_j(\cdot). \tag{1}$$

Feature fusion-based multi-modal networks. Many existing multi-modal FL solutions [17] can only work for result-level fusion models, which cannot take advantage of low-level correlations between the modalities. Here we consider a feature-level multi-modal fusion network, which merges the feature representations extracted by unimodal encoder networks of different modalities before making the prediction. As discussed in 2, feature-level fusion can exploit the correlation and interactions between features of different modalities. Moreover, it is compatible with more deep learning-based multi-modal fusion approaches as it can re-use the pre-trained backbone model from each modality [11].

Specifically, on a multi-modal node $c_k(m) (1 \leq k \leq N_m)$ that has M_k data modalities, the data of each modality will be separately fed into the unimodal feature encoder $\{f_{enc_i}(\cdot) | \forall i \in \mathcal{M}_k\}$ to generate M_k representation vectors:

$$\mathbf{h}_i = f_{enc_i}(\mathbf{x}_i), i \in \mathcal{M}_k, \tag{2}$$

where $\mathbf{h}_i \in \mathbb{R}^{D_i}$ is the hidden feature of the i_{th} sensor modality extracted by $f_{enc_i}(\cdot)$. Here the unimodal feature encoders can be

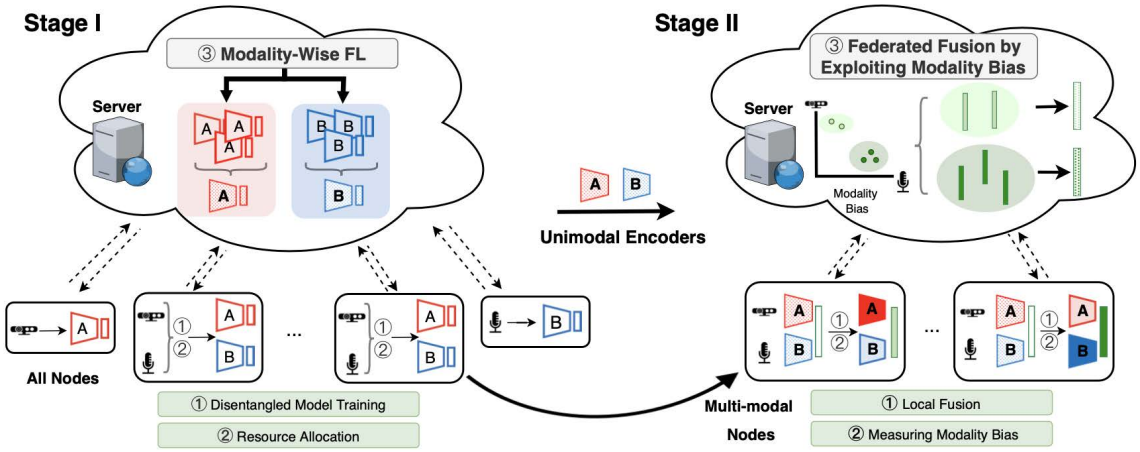


Figure 5: Harmony consists of two stages, i.e., modality-wise federated learning among all nodes, and federated fusion among multi-modal nodes by exploiting the modality biases, respectively.

any off-the-shelf deep learning models (e.g., convolutional neural network [40] or recurrent neural network [64]) depending on the sensor modalities. This implies that our framework is general and can be adopted in various applications.

Then the extracted unimodal features will be fused using direct feature concatenation [62] or attention-based concatenation [35] to combine the complementary information of different modalities. The fused features will be input into several fully connected layers to make the prediction. Then the fusion-based classifier in the multi-modal network $g(\cdot)$ can be expressed as:

$$y_{pred} = g(\mathbf{h}_i | \forall i \in \mathcal{M}_k), \tag{3}$$

Then the feature fusion-based multi-modal network of the node $c_k(m)$ can be expressed by:

$$\Phi_k(m) = \{f_{enc_i}(\cdot) | \forall i \in \mathcal{M}_k\} \cup g(\cdot). \tag{4}$$

Although the multi-modal and single-modal networks have different model architectures, the modality-specific feature encoders share the same structure among the nodes that contain the same data modalities. Therefore, the server in multi-modal FL systems can merge the information of single-modal and multi-modal nodes by properly aggregating the unimodal feature encoders.

5.2 System Architecture

The design of Harmony is motivated by the key insights from Section 4 that, directly aggregating feature encoders trained by multi-modal and single-modal networks will result in model performance degradation on both sides. Therefore, our key idea is to *disentangle the multi-modal training* into a novel two-stage framework, namely *modality-wise federated learning* and *federated fusion learning*. Figure 5 shows the overall system architecture of Harmony. In the following, we refer to the collection of nodes that perform the same single-modal network training as a *unimodal FL subsystem*.

In *modality-wise federated learning*, the multi-modal nodes train multiple single-modal networks rather than a single multi-modal fusion network. Therefore, this stage runs multiple *unimodal FL subsystems* in parallel, where all nodes consistently learn unimodal information. Such a scheme naturally reduces the model divergence within a specific FL subsystem and improves the model performance of single-modal nodes. This stage will also converge faster than conventional multi-modal FL systems, since each unimodal FL subsystem will involve fewer nodes with homogeneous data modalities. As the multi-modal nodes will participate in multiple unimodal FL subsystems, we propose to *dynamically optimize the resource utilization* of multi-modal nodes to different single-model training tasks. In particular, the multi-modal nodes will allocate more computing resources to the modalities whose unimodal FL subsystems take longer to aggregate models in a single training round. Such a scheme will balance the delays of different *unimodal FL subsystems* and hence speed up the overall system convergence.

In *federated fusion learning*, only multi-modal nodes will collaboratively train the classifier layers, which is challenging when the nodes have non-i.i.d data distribution. As shown in Section 4, the model performance of different multi-modal nodes may rely on different modalities. We design a novel federated fusion mechanism that measures and exploits the *modality biases* of different multi-modal nodes in the aggregation of classifiers. Specifically, all multi-modal nodes will initialize the feature encoders using model weights trained in modality-wise FL and then fine-tune them. Therefore, the pre-trained encoders serve as a benchmark to quantify the discrepancy of unimodal encoders during local fusion, which essentially reflects the modality bias of local data. This is a unique advantage brought by our two-stage training framework and can be used to assist the aggregation of classifiers. In particular, the server will cluster the nodes according to their modality biases and aggregates the classifiers in each cluster, where the nodes with similar modality biases will be in a cluster. Based on the pre-trained feature encoders, federated fusion learning will converge fast and incur low system overhead.

After the two-stage training, the server in Harmony will send both multi-modal and single-modal models to all nodes. Therefore, the nodes can select either multi-modal or single-modal models during inference, according to the modality of their local data over time. This feature allows Harmony to adapt to runtime dynamics such as sensor failures. For instance, in Alzheimer's patients monitoring, when the mmWave radar fails (e.g., due to power surges or

unstable sensor connection), the nodes can still utilize the depth sensor for behavior analysis based on the single-modal network.

6 DESIGN OF HARMONY

The design of Harmony is motivated by the key observation that, directly aggregating the unimodal encoders trained by multi-modal fusion networks and single-modal networks will result in model performance degradation on both sides. Therefore, we propose to disentangle the training of multi-modal networks in a novel two-stage framework, namely *modality-wise federated learning* and *federated fusion learning*.

6.1 Modality-Wise Federated Learning

In modality-wise FL, all nodes in the system will collaboratively train single-modal networks of different data modalities. The feature encoders of the trained single-modal networks can then be reused by multi-modal nodes in federated fusion learning. We will first introduce the federated learning framework and then present how to speed up the system convergence through resource allocation of the multi-modal nodes.

6.1.1 Disentangled Model Training. As shown in Section 4, directly aggregating the feature encoders trained by multi-modal and single-modal learning will result in model performance degradation on both sides. Therefore, instead of performing local fusion on the multi-modal nodes before model aggregation with single-modal nodes, we propose to disentangle the training of feature encoders $\{f_{enc_i}(\cdot)|\forall i \in \{1, ..., M\}\}$ and the classifier $g(\cdot)$ into two stages.

As shown in Figure 5, in modality-wise FL, the multi-modal nodes will train multiple single-modal networks rather than multi-modal fusion networks. For example, for a multi-modal node $c_k(m)(1 \leq k \leq N_m)$ with M_k different modalities, it will train total M_k different single-modal networks:

$$\tilde{\Phi}_k(\cdot) = \{\Phi_k(s_i)|\forall i \in \mathcal{M}_k\}, \tag{5}$$

Here $\Phi_k(s_i) = f_{enc_i}(\cdot) \cup g_i(\cdot)$ denotes the single-modal network of modality s_i. Moreover, a single-modality node $c_k(s)$ will still train a single-modal network $\Phi_k(s)$ with the architecture shown in Eq. (1).

Through disentangled model training, the multi-modal nodes can capture more modality-specific useful information, while some of it would be lost during multi-modal fusion learning. Moreover, multi-modal networks have a larger size of model parameters and are likely to overfit on local training data [55]. In contrast, the feature encoders trained by different single-modal networks will have better generalization ability, which can be reused in federated fusion learning. Finally, such a scheme is more robust for practical scenarios where the data modalities on nodes change over time due to dynamic sensor failure. In this case, the multi-modal nodes in Harmony can still leverage data with partial modalities to train single-modal networks.

6.1.2 Parallel Unimodal Federated Learning. After disentangling the training of multi-model models, all nodes will train and upload single-modal networks in modality-wise FL. Therefore, there will be multiple unimodal FL subsystems running in parallel. At the $(r+1)$-th communication round, the procedures of node update and server update are as follows.

- Node Update: The node c_k will parallelly optimize (e.g., using gradient descent methods) the model weight of M_k single-modal networks based on its local data $(\{\mathbf{x}_i|\forall i \in \mathcal{M}_k\}, y)$.

$$\Phi_k^{r+1}(s_i) \leftarrow \mathbf{SGD}(\Phi_k^r(s_i), (\mathbf{x}(i), \mathbf{y})), i \in \mathcal{M}_k. \tag{6}$$

- Server Update: The server will run M different threads for handling the model aggregation of different unimodal FL subsystems. For modality $j \in \{1, 2, ..., M\}$, if the model weights of all nodes (where there are N_j nodes that have the data of modality j) have arrived at the server, the server will perform the model aggregation as:

$$\overline{\Phi}^{r+1}(s_j) = UniFL(\Phi_1^{r+1}(s_j), ..., \Phi_{N_i}^{r+1}(s_j)). \tag{7}$$

Here the aggregation approach "UniFL(·)" can be any existing FL algorithms that aim to generate a single global model (e.g., FedAvg [38] or FedProx [30]). Therefore, the modality-wise FL of Harmony can be integrated with many state-of-the-art FL algorithms to further improve the model performance in the presence of non-i.i.d. data distribution among nodes.

Through modality-wise federated learning, the feature encoders are trained to consistently extract unimodal information. This will naturally harness the modality heterogeneity in a multi-modal FL system, since it eliminates the negative impact of aggregating the encoders trained by single-modal and multi-modal learning. Moreover, the unimodal encoders are trained on all distributed data of the corresponding modalities, which will have a better generalization ability. Finally, this stage will converge faster than conventional multi-modal FL systems. This is because each unimodal FL subsystem will involve fewer nodes with homogeneous data modalities, significantly reducing the global round completion time.

6.1.3 Balance-Aware Resource Allocation. In this section, we design a novel resource allocation mechanism for multi-modal nodes that balances the training delays of different modalities and nodes to reduce the overall system latency of modality-wise FL.

Basically, the overall delay of a FL system consists of three parts: the computing time of on-device model training, the communication time of model transmission, and the waiting time of synchronizing all nodes for model aggregation. In multi-modal FL systems, the models on nodes with different modalities usually have various convergence speeds and model sizes [55], resulting in highly heterogeneous computing and communication delays. Moreover, resource availability usually suffers significant dynamics in mobile and edge systems, making it challenging to reduce the overall system delay.

In Harmony, the multi-modal nodes train multiple single-modal networks in parallel and will likely be the bottleneck of the system convergence. However, thanks to the disentangled model training, the multi-modal nodes can coordinate the convergence of different unimodal FL subsystems. For example, as shown in Figure 6, the multi-modal nodes run two processes for single-modal training of audio and depth data, respectively. Due to different model sizes and input data dimensions, the processes of the two modalities take different times, e.g., 15s for depth and 7.5s for audio on Node 1. Moreover, the training and communication delays of Node 1 and Node 2 are also different due to diverse resource availabilities. As a result, the unimodal FL subsystems of depth and audio have imbalanced convergence latencies. Such imbalance can lead to higher

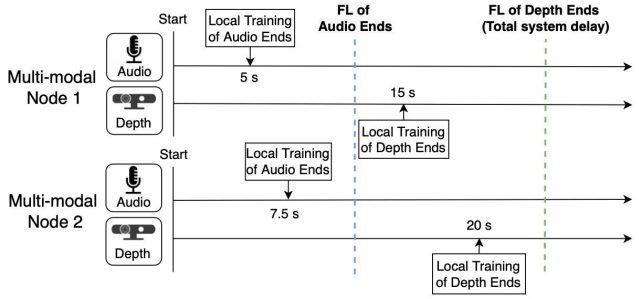


Figure 6: An example of imbalanced delays among local single-modal training processes and global unimodal FL subsystems. The delay imbalance prolongs the overall system convergence.

overall system latency, which is determined by the slowest subsystem (i.e., depth images). Therefore, in order to reduce the overall system delay, our key idea is to optimize the resource allocation of different multi-modal nodes to tackle the imbalanced training delays among different nodes and modalities.

Resource allocation for multi-modal nodes. Motivated by the aforementioned example, our objective is to dynamically compute a resource ratio vector that accounts for imbalanced delays, thereby enabling balance-aware resource allocation on multi-modal nodes to reduce the overall system delay. This is achieved by factoring in two key considerations when calculating the ratio vector for each multi-modal node: the convergence performance of different unimodal FL subsystems (inter-modality balance), and the local resources available to each node (intra-modality node performance).

To quantify the convergence performance, in each round, the server will measure the round completion time of each FL subsystem T^i for modality $i \leq (1 \leq i \leq M)$ and of in the subsystem (e.g., t_k^i for the k-th node in modality i's subsystem). Based on the delay measurements, the server then calculates the resource ratio vectors $\beta_{\mathbf{k}} = [\beta_k^i, ..., \beta_k^M], 1 \leq k \leq N_m$ for single-modal training tasks on different multi-modal nodes. Specifically, on a multi-modal node $c_k(m)$, the resource allocated to modality i's single-modal training should account for the following two aspects: (1) Inter-modality balance: the rank of round completion time of modality i's FL subsystem among all modalities; (2) Intra-modality node performance: the rank of node $c_k(m)$'s round completion time among all nodes with modality i. Thus, the ratio value β_k can be calculated by:

$$\beta_k^i = \frac{T^i}{\sum_{i=1}^{M} T^i} \cdot \frac{t_k^i}{\sum_{k=1}^{N_i} t_k^i / N_i}, i = 1, ..., M, \tag{8}$$

where the left part quantifies the convergence performance of modality i over all subsystems, and the right part compares the model updating efficiency of node $c_k(m)$ with the average of all nodes. Then the server will normalize the ratio vector β_k and send it to the corresponding multi-modal node $c_k(m)$. Therefore, the resource ratio vectors are customized for nodes with different local resources and training tasks. Through reallocating the computing resources to different unimodal training tasks, the convergence delays of the different unimodal FL subsystems will be more balanced, thus reducing the overall system delay.

Local execution with assigned ratio. After receiving the ratio vector, each node then re-allocates its local computing resource proportionally among all local modalities. Our ratio-based resource allocation scheme can be implemented by using various existing resource scheduling algorithms. For example, based on the ratio vector, a multi-modal node can assign the numbers of CPU/GPU cores or set the time slices to different single-modal training tasks. In our implementation, we use priority-based scheduling with time slicing, where a multi-modal node will set the priority and time slices (i.e., execution time) for different tasks according to the received resource ratios. Moreover, if a modality finishes one round of local model training earlier, its training task will sleep to free its remaining resources to the single-model training of other modalities, thus fully utilizing the computing resources on devices. We now revisit the example in Figure 6 to show how our design reduces the overall system delay. By calculating the resource ratio vectors with the measured delays and Equation (8), the two multi-modal nodes (especially Node 2) will allocate more computing resources to the single-modal training task of depth images. As a result, the convergence delays of the two unimodal FL subsystems will be more similar, thus reducing the overall system delay.

6.2 Exploiting Modality Bias in Federated Fusion

In federated fusion learning, the multi-modal nodes collaboratively train the classifiers that fuse the unimodal features and make predictions based on the fused features. However, aggregating the classifiers of multi-modal nodes is challenging in the presence of non-i.i.d data distributions. We design a novel federated fusion mechanism that exploits the modality biases of different multi-modal nodes to improve the model performance.

6.2.1 Measuring Modality Bias via Encoder Discrepancy. As introduced in Section 4, the multi-modal networks of different nodes may show substantial bias toward different modalities. We propose to measure and leverage such modality biases in different multi-modal networks to address the data non-i.i.d problem.

The key idea is to quantify the modality biases of different multi-modal nodes using the discrepancy of their unimodal encoders during local fusion. In particular, all multi-modal nodes will initialize the feature encoders using model weights trained in modality-wise FL and then fine-tune them. In other words, the pre-trained unimodal encoders are the starting points of multi-modal learning, which serves as a benchmark for all nodes in federated fusion. As a result, the discrepancy of unimodal encoders during local fusion essentially reflects the data modality biases of different nodes.

Specifically, at the r-th communication round in federated fusion learning, the multi-modal nodes will calculate the cosine distance between the model weights of current encoders (after local fusion learning) and the initial benchmark encoders. For modality i in node $c_k(m)$, the encoder discrepancy is calculated by:

$$d_k^r(i) = dis(f_{k,enc_i}^r(\cdot), f_{enc_i}^0(\cdot)). \tag{9}$$

Here $dis(\cdot)$ measures the cosine distance of two weight vectors. Then the node $c_k(m)$ will send the encoder discrepancy vector $\mathbf{d}_k^r = [d_k^r(i), ..., d_k^r(M_k)]$ to the server.

6.2.2 Cluster-based Fusion Aggregation. After receiving the encoder discrepancy vectors and model weights of classifiers from all multi-modal nodes, the server will cluster the nodes according to their modality biases and aggregate the classifier layers with each cluster. In contrast to uni-modal FL studies that cluster nodes based on the entire model parameters [43, 44], we utilize the modality bias of multi-modal nodes as the clustering metric for the following reasons. First, the data modality biases (i.e., the difference in uni-modal encoders during local fusion) essentially reflect the non-i.i.d. data distributions of different multi-modal nodes, making it easier for clustering than using the entire multi-modal networks [43]. Additionally, this approach significantly reduces the communication delay since the modality biases can be calculated locally such that only the classifier needs to be transmitted.

Specifically, the server will first normalize the encoder discrepancy value of each modality among all nodes. For the modality i of the k-th multi-modal node:

$$d_k^r(i) = \frac{d_k^r(i)}{\max\{d_1^r(i), ..., d_{N_q}^r(i)\}}, k = 1, ..., N_q. \quad (10)$$

Then each node will have a normalized vector of encoder discrepancy $\mathbf{d}_k^r \in \mathbb{R}^{M_q}$. According to the normalized encoder discrepancy vectors $\{\mathbf{d}_1^r, ..., \mathbf{d}_{N_q}^r\}$, the server will group the N_q nodes to K_q different clusters using K-means [18]. Figure 7 visualizes an example of the normalized encoder discrepancy values among multi-modal nodes on the MHAD dataset [41], where the accelerometer (Acc) and skeleton data are used. It is shown that the nodes form different clusters on the 2D space of the encoder discrepancy of Acc and skeleton. When setting the number of clusters in K-means as three, we can easily obtain the clustering result of the six multi-modal nodes as [[0,1,4,5], [2], [3]].

The number of clusters K in the K-means clustering can be determined based on some prior knowledge of the major types of nodes in FL. For example, in the evaluation of Alzheimer's Disease monitoring, we set $K = 3$ as there are mainly three groups of users (i.e., with AD, with mild cognitive impairment, and cognitively normal). On the other hand, if we do not have such prior knowledge, K can be decided based on numerical analysis results. Specifically, we can gather the encoder discrepancy vector of all nodes $\{\mathbf{d}_i^r \in \mathbb{R}^{M_q}, \forall i = 1, ..., N_q\}$ as a matrix $\mathbf{D} \in \mathbb{R}^{M_q \times N_q}$ and perform singular value decomposition [19] on the matrix $\mathbf{D}$. Then K can be set as the number of dominant singular values of $\mathbf{D}$. For example, we can set $K = 3$ if the singular values are [100, 50, 30, 1, 0.5, 0.1, 0].

Next, the server will aggregate the classifiers of multi-modal nodes within the same cluster. Federated fusion among the same group can leverage existing FL algorithms that train a single global model, where we use FedAvg [38] unless otherwise specified.

Reducing Communication Overhead. Federated fusion learning will incur only a small system overhead due to the following reasons. First, based on the unimodal feature encoders trained in modality-wise FL, the local training of multi-modal networks will converge fast. Second, the multi-modal nodes will only transmit the model updates of classifier layers and encoder discrepancy vectors to the server, resulting in small communication latency. Third, the number of involved nodes (only multi-modal nodes) in federated fusion learning is much smaller than in the original multi-modal FL system, significantly reducing the overall convergence latency.

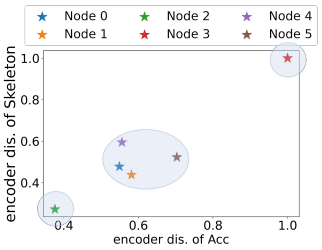


Figure 7: Visualization of encoder discrepancy vectors of multi-modal nodes. The nodes are grouped into three clusters based on the encoder discrepancy.

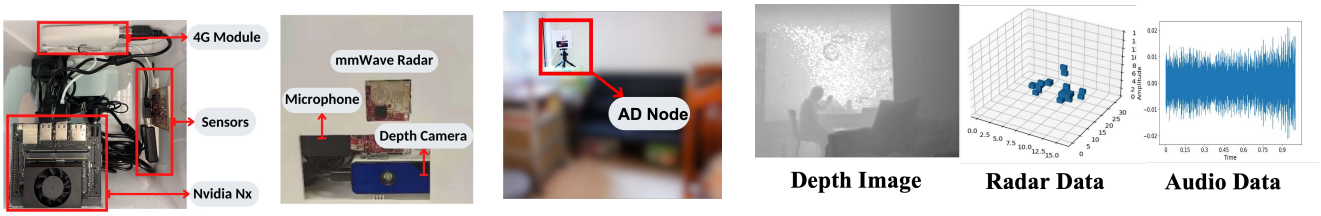
7 REAL-WORLD TESTBED EVALUATION

7.1 Real-World FL Testbed

We implemented Harmony on a real-world multi-modal sensor system deployed in homes of elderly subjects[2]. The system is designed to classify multi-dimension digital biomarkers (e.g., Activities of Daily Living, Behavioral and Psychological Symptoms of Dementia [14], motor functions, and cognition) for early diagnosis and intervention of Alzheimer's Disease. Figure 8 shows the overview of our multi-modal testbed.

The testbed consists of 16 sensor nodes installed in participants' homes and a central server located in our lab. We developed a compact hardware system, which incorporates three privacy-preserving sensors (a depth camera, a mmWave radar, and a microphone), an NVIDIA Xavier NX single-board edge computer [4] with 1TB external NVMe SSD, and a 4G cellular interface to communicate with the lab server. The nodes can collect and store multi-modal data, train deep learning models locally, and communicate with the server for federated learning. We choose the three sensor modalities to collectively capture a wide range of biomarkers while preserving the users' privacy. In particular, the depth camera can detect context-aware activities like cleaning living areas and moving in/out of chairs, but cannot reveal sensitive personal information like faces. The mmWave radar can capture motion-related activities like walking, standing, and sleeping. The ambient microphones can help detect acoustic-related activities like eating, drinking, and phone calls without recording raw acoustic data. The NVIDIA Xavier NX edge device (6-core ARM CPU and 84-core GPU, 16GB Memory) runs Ubuntu 18.04, and the lab server (Intel Core i5-12400 CPU, RTX3060 GPU, 32 G Memory) runs Ubuntu 22.04. The sensor data sampling and machine learning models are implemented using Python 3. In order to capture the main living area and reduce the domain gap of different subjects' data, we put the node at a height of 1.5m-1.8m in the living room (typically on the shelf or cabinet around the sofa), and used a tripod to adjust the height and angle of the box. The area of the subjects' living room is $10m^2$-$25m^2$. Harmony runs on these nodes continuously for four weeks.

[2]All the data collection was approved by IRB and the Clinical Research Ethics Committee of the authors' institution.



(a) Our multi-sensor hardware prototype. (b) Home installations. (c) Examples of recorded multi-modal data.

Figure 8: Our real-world multi-modal sensor testbed for Alzheimer's Disease monitoring. The nodes incorporating three sensor modalities (depth, mmWave radar, and audio) are deployed in the homes of 16 elderly subjects.

7.2 Human Subjects and Data

The deployed system detects behaviors that are shown to be highly related to Alzheimer's Disease from the medical literature. A total of 16 elder subjects (eight females and eight males, 62-80 years old) have participated in our study, including six with Alzheimer's Disease, six with mild cognitive impairment (MCI) and four cognitively normal subjects. The subjects live alone or with their families, and follow their usual daily routines during the period of four continuous weeks. The activity distributions of different subjects are highly heterogeneous. First, AD patients usually have less mobility and tend to have a sedentary lifestyle compared with cognitively normal subjects. For example, AD subjects spend more time on basic living activities like sitting, standing, and sleeping. Second, cognitively normal and MCI subjects exhibit more diverse activities than AD subjects. For example, the average number of occurred activities in AD, MCI, and cognitively normal subjects is 6.85, 8.75, and 9, respectively, which shows the decline in cognitive and functional ability during the progression of AD [33].

To reduce the energy consumption and storage overhead, the sensors will go to sleep when there is little user activity. In total, the 16 nodes collected 10,752 hours of multi-modal sensor data, with a total size of 16TB. The multi-modal data are synchronized using the system clock and annotated using depth videos. This is because depth video is easier for humans to annotate and has a smaller sensing range compared to mmWave radar and microphone [58, 59]. We removed the data samples where the subjects are out of the range of the depth camera in the evaluation.Finally, we focused on the data recorded from 6 am to 12 am of one day, and obtained about 96 hours of labeled multi-modal data (some with only partial modalities). The sampling rates of the depth camera, mmWave radar, and microphone are 15 Hz, 20 Hz, and 44,100 Hz, respectively. We split the sensor data into 2-second samples and converted them into a fixed dimension, i.e., [16,112,112], [20,2,16,32,16], and [20,87] for depth, radar, and audio data, respectively. We discarded the classes that have very limited samples from most of the subjects, and the remaining data fall into 11 classes, including cleaning the living area, taking medication, using mobile phones, writing, sitting, standing, moving in/out of chair/bed, walking, sleeping, eating, and drinking. These activities are shown to be highly related to Alzheimer's Disease in the medical literature [9, 36, 48] and can be detected in home environments. The duration and frequency of these activities can be used as potential digital biomarkers for diagnostic analysis of AD [2, 7, 8]. Finally, the number of labeled samples on the nodes is in [1141, 6498], and the total number of labeled samples is about 60,000.

	Sensor combination
Set 1	2A, 2D, 2R, 10(A,D,R)
Set 2	2(A,D), 2(D,R), 2(A,R), 10(A,D,R)
Set 3	1A, 1D, 1R, 2(A,D), 2(D,R), 2(A,R), 7(A,D,R)

Table 1: Selected sensor combinations on 16 nodes. A, D, R denotes Audio, Depth, Radar, respectively, and 7(A,D,R) means seven nodes having three modalities.

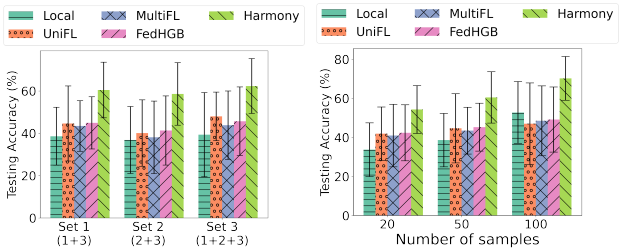
7.3 Results on Real-World Testbed

7.3.1 Evaluation metrics and configurations. We evaluate the system accuracy of behavior recognition and the wall clock time for the whole FL system to achieve convergence. Our baselines include local learning, UniFL, MultiFL (introduced in Section 4), and FedHGB [17]. In FedHGB, the nodes use a validation dataset to measure the overfitting-to-generalization rate, according to which the server computes the aggregation weights among modalities and nodes. The reasons why we choose these baselines are as follows. First, existing multi-modal FL approaches such as MultiFL [60, 67] used as our baseline that can work with feature-level fusion are largely based on Fedavg. Second, although FedHGB can only be applied for result-level fusion, it is a state-of-the-art approach that aims to address the non-i.i.d problem under modality heterogeneity in multi-modal FL, thus serving as a strong baseline.

During the modality-wise FL of Harmony, we adopt different processes on multi-modal nodes for different single-modal training tasks. The server calculates the resource ratio vectors for different nodes in each communication round according to our design in Section 6.1.3 and sends them to the nodes. Then the multi-modal nodes will dynamically allocate CPU and GPU time slices to different processes using priority-based time slicing scheduling [45].

7.3.2 System Accuracy. In this section, we investigate the characteristics of the real-world multi-modal data collected by our testbed and evaluate the system accuracy of Harmony.

Dynamics of sensor modalities. In our testbed, the sensors may stop data recording occasionally due to the system dynamics, such as power surges or unstable sensor connections. We use the *systemd* service [5] of Linux to restart the sensors in case of sensor faults. As a result, the number of working sensors changes over time as well as across different nodes. We sample the status of sensors every hour during the four weeks, and then average the results of 16 nodes. On a specific node, the mean probability of having three, two, one and zero working sensors are 95.82%, 3.41%, 0.37% and 0.02%, respectively. Thanks to the disentangled multi-modal



(a) Different modality sets. (b) Different amounts of data.

Figure 9: Accuracy performance on real-world multi-modal data. Harmony outperforms by 20% in mean accuracy over the baselines under various settings.

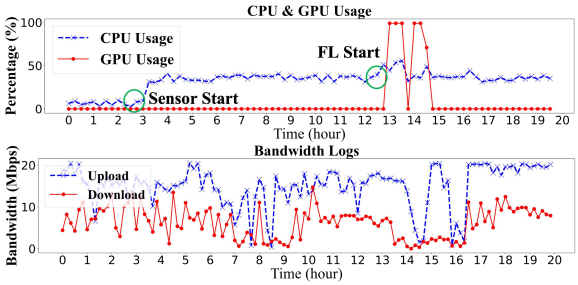


Figure 10: Dynamics of resources on the nodes.

training, Harmony can leverage the data of partial modalities on multi-modal nodes, which is robust to runtime system dynamics such as sensor failures.

Accuracy of different modality combinations. Due to the system dynamics, the data modalities currently available on a given node may have various combinations. We select three representative sensor combinations (shown in Table 1) for the 16 nodes and evaluate the accuracy of Harmony under these settings. Figure 9a shows the accuracy when each node only has 50 labeled training samples. Harmony delivers significant accuracy improvement over the state-of-the-art baselines across different settings of sensor combinations. For example, in Set 3, Harmony outperforms 22.96%, 14.33%, 18.48%, and 16.65% over local learning, UniFL, MultiFL and FedHGB, respectively.

Performance with different amounts of local data. We further evaluate the performance of Harmony with different amounts of local training data. As shown in Figure 9b, when there are more local training samples, generally, the accuracy of all approaches increases. However, due to the modality and data heterogeneity, UniFL, MultiFL and FedHGB only have little accuracy improvement or even perform worse than local learning. Harmony consistently outperforms the baselines with different amounts of training data, by effectively exploiting the heterogeneous data across the nodes. For example, Harmony achieves over 70% accuracy with only 100 training samples on nodes.

7.3.3 System Overhead. We then evaluate the system overhead of Harmony on our testbed under various resource dynamics.

Dynamic compute and communication resource availability. A key challenge addressed in the design of Harmony is the significant resource dynamics of nodes in real-world FL systems.

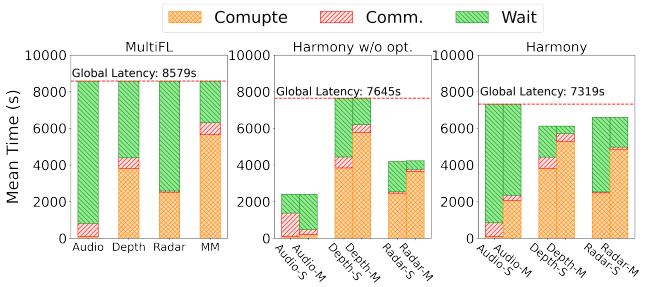


Figure 11: System latency of different FL solutions.

Figure 10 plots the CPU/GPU usage and bandwidth of a multi-modal node over 20 continuous hours. After the multi-modal sensor data recording starts, the available computing resources change over time. When the node starts to run real-time federated learning, the GPU usage nearly approaches 100%. Moreover, over the 20 hours, both the uplink and downlink bandwidth of the 4G LTE networks fluctuate in a significant dynamic range (e.g., 0-20 Mbps).

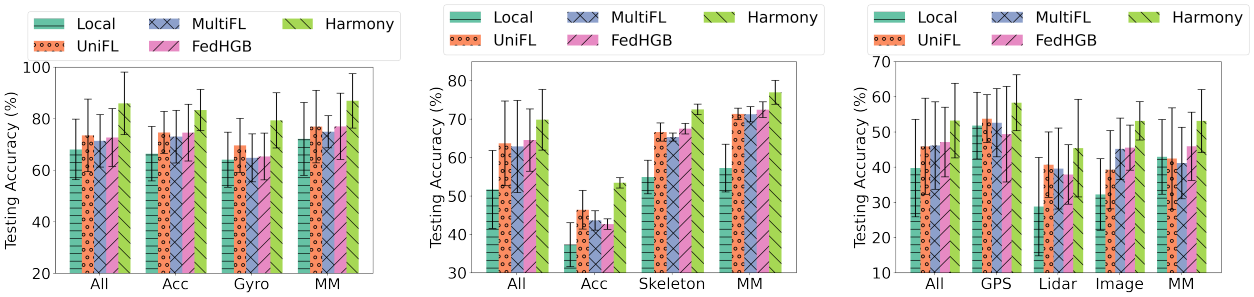
System training latency. Figure 11 compares the mean training latency of different types of nodes in MultiFL and the modality-wise FL of Harmony. In MultiFL (left figure), the multi-modal nodes (MM) have the highest computing delay. Consequently, the single-modal nodes (Audio, Depth, Radar) incur a long waiting time for global model aggregation. In Harmony without resource allocation in Section 6.1.3 (middle figure), the multi-modal node runs the process of single-modal training of audio, depth, and radar in a round-robin manner. As a result, the audio model with the smallest training workload finishes unimodal FL first, followed by the radar and depth model. In Harmony (right figure), thanks to the balance-aware resource allocation design, the multi-modal nodes allocate more computing resources (i.e., the highest priority and largest ratio of time slice) to depth's unimodal training, which converges first. Then the freed resources will be used by the radar and audio model, which converges subsequently. As a result, the modality-wise FL of Harmony converges faster than MultiFL. Furthermore, based on the pre-trained unimodal encoders, federated fusion learning in Harmony only takes 235s to converge, which is a fraction of 3.2% delay of modality-wise FL. Therefore, the two-stage training framework and resource allocation design of Harmony together reduce the overall latency in heterogeneous multi-modal FL systems.

8 EVALUATION ON PUBLIC DATASETS

8.1 Datasets

USC dataset [66]. This dataset comprises data of a 3-axis accelerator and 3-axis gyroscope from 14 users performing 12 activities. The sampling rate of the two sensors is 100 Hz. We choose a 2-second time window that generates a 600-dimensional vector for data of each modality. We set the data of different subjects to different nodes, where eight with data of both modalities, three with only accelerometer data, and three with only gyroscope data.

MHAD dataset [41]. This dataset contains data of 11 human actions collected from 12 subjects. Due to the small amount of data from each subject, we use the 3-axis accelerometer and skeleton data with a relatively low dimension to avoid model overfitting. We resample the data of both modalities to 30 Hz, and each frame



(a) USC (b) MHAD (c) FLASH

Figure 12: Comparison of accuracy performance on different multi-modal datasets. Harmony consistently outperforms the state-of-the-art baselines for nodes with different data modalities.

Dataset	Modality	Class	Node	Samples
USC	Acc, Gyro	12	14	38312
MHAD	Acc, Skeleton	11	12	3956
FLASH	GPS, LiDar, Camera	64	210	32923

Table 2: Summary of the three multi-modal datasets.

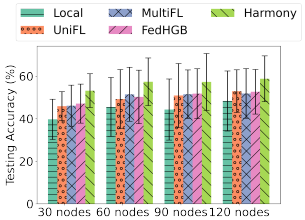


Figure 13: Accuracy with various numbers of nodes.

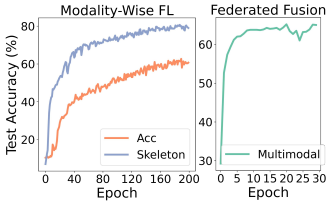


Figure 14: Convergence performance.

of multi-modal data consists of 3-D accelerometer data and 35×3-dimensional skeletal points. We choose a sliding time window of 2s to generate around 330 samples per subject. We assign six nodes with data from both modalities, three with only accelerometer data and three with only skeleton data.

FLASH dataset [47]. This dataset comprises data of GPS, LiDAR, and cameras collected using autonomous cars at 10 Hz. The task is to select the high-band sector for mmWave beamforming in mobile V2X communication scenarios. Each sample contains a 64-dimensional RF ground truth and synchronized multi-modal sensor data, with the dimension [1,2], [20,20,20], and [3,360,640] for GPS, LiDAR, and image, respectively. To evaluate the scalability of Harmony, we assign the data from different vehicles and scenarios to different nodes and divide the whole dataset into 210 nodes. The number of nodes is proportionally set as 4:2:2:2 for the nodes with multi-modal, GPS, Lidar and image data, respectively.

8.2 Implementation

To evaluate the scalability of Harmony, we set up more nodes on a computing cluster consisting of eight powerful but heterogeneous machines. Each machine contains four GPU cores and 16/32 CPU cores. We assign different CPU cores to nodes on the same machine and let each GPU run multiple nodes to incorporate up to 210 nodes. We use CNN layers to extract deep features, RNN layers to capture the time-series properties, and two fully-connected layers for the classifier. For the unimodal feature encoders, we adopt 2D-CNN for the inertial/GPS data, 3D-CNN for the skeleton, Lidar and image data. The learning rate and batch size are set as 0.001 and 16 for Harmony and the baselines. Each experiment is repeated five times.

8.3 Overall Performance

In this section, our evaluation focuses on three aspects of Harmony, including performance on different datasets, scalability and convergence performance.

Accuracy on different datasets. Figure 12 plots the accuracy of different types of nodes on the three public multi-modal datasets. First, Harmony shows significant accuracy improvement over the baselines on nodes with heterogeneous data modalities, e.g., outperforms by 18.67%, 5.76%, 6.97%, and 5.3% over local learning, UniFL, MultiFL, and FedHGB on the MHAD dataset, respectively. Second, compared with UniFL or local learning, the single-modal nodes in MultiFL/FedHGB generally suffer vivid accuracy drops, as their models are aggregated with multi-modal networks. Harmony harnesses the negative impacts through disentangled multi-modal training. For example, on the USC dataset, Harmony improves the mean accuracy of Acc, Gyro, and multi-modal nodes by 15.34%, 14.34%, and 19% over MultiFL.

Scalability. Figure 13 shows the accuracy performance of Harmony with different numbers of nodes (i.e., 30, 60, 90, 120), using data from the FLASH dataset. Generally, when the number of nodes increases, the variance of model accuracy among nodes in local learning increases, which shows the heterogeneity of nodes' data. Moreover, the accuracy performance of FL-based approaches increases with more nodes. Harmony consistently outperforms local learning, UniFL, MultiFL, and FedHGB in model accuracy in all configurations, demonstrating its scalability.

Convergence Performance Figure 14 shows the convergence performance of Harmony over the training epochs on the MHAD dataset. In modality-wise FL, the unimodal FL subsystems of Acc and Skeleton converge at different speeds. Federated fusion learning in Harmony converges fast and only needs about 10 to 15 epochs of local training to achieve the highest accuracy. This is because federated fusion learning is based on the feature encoders pre-trained on large amounts of distributed data in modality-wise FL.

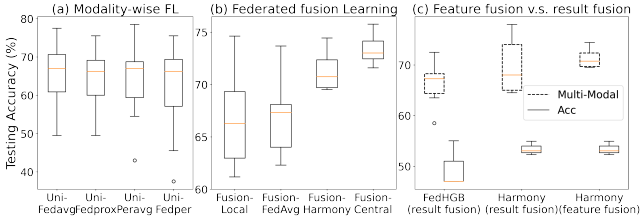


Figure 15: Ablation study of Harmony.

8.4 Understanding Harmony's Performance

In this section, we perform the ablation study to understand the effectiveness of Harmony. We show the results of the MHAD dataset, and the results on the other two datasets are similar.

Integrating existing FL algorithms. We first compare the model accuracy of all nodes when the modality-wise FL of Harmony runs different personalized FL algorithms, including FedProx [30], PerAvg [20] and FedPer [10]. The results are shown in Figure 15(a). We observe that integrating personalized FL algorithms in modality-wise FL does not always improve the model accuracy of all nodes. For example, running Fedper in the modality-wise FL (Uni-Fedper) introduces an accuracy drop to some nodes. This is because the multi-modal nodes reuse the feature encoders trained in modality-wise FL, where more personalized encoders may not have good adaptation performance in federated fusion learning.

Effectiveness of federated fusion learning. To show the effectiveness of our federated fusion mechanism, we compare the model accuracy of multi-modal nodes when their classifiers are trained by different schemes, including local fusion, FedAvg, federated fusion learning of Harmony, and centralized fusion. We initialize the modal weights of encoders using the one trained in modality-wise FL of Harmony. The results are shown in Figure 15(b). Compared with local fusion and FedAvg, Harmony improves the model accuracy of most multi-modal nodes, and its mean accuracy even approaches the performance of centralized fusion learning.

Comparing feature-level and result-level fusion. To understand the performance gain of our approach, we compare the accuracy of FedHGB (only applicable to result fusion) [17], Harmony with result fusion, and Harmony with feature fusion. Figure 15(c) shows the model accuracy of multi-modal and single-modal (with only accelerator data) nodes, when they run the above three FL methods. First, when the multi-modal nodes train result-level fusion models of different modalities, Harmony can improve model accuracy of most nodes compared with FedHGB, but does not need a validation dataset. Second, besides the result-level fusion, Harmony can also be applied to feature-level fusion models, further improving the accuracy of multi-modal nodes. Third, the single-modal nodes (Acc) in Harmony also have better accuracy performance than FedHGB, thanks to the modality-wise FL in Harmony.

9 DISCUSSION

Extension to semi-supervised learning. There is usually limited labeled data in real-world applications like Alzheimer's Disease monitoring, which is a key motivation for adopting FL. Although Harmony works reasonably well with limited data, e.g., achieving over 70% accuracy with only 100 training samples in real-world AD monitoring, its performance can be further improved by adopting semi-supervised learning approaches. Specifically, the modality-wise FL in Harmony aims to train feature encoders, which does not rely on labeled data, and thus can be easily extended with unsupervised learning approaches that train feature encoders with large amounts of unlabeled data. As a result, the classifiers of nodes can be trained in federated fusion learning with only limited labeled data. Previous studies on semi-supervised learning have shown that the model performance can be augmented by efficiently learning from the unlabeled multi-modal [42] or unimodal [23, 61] data.

Impact of runtime sensor dynamics. Although Harmony allows the nodes to select either multi-modal or unimodal models during inference according to their local data modality at runtime, the system performance may vary during the switch of the inference model. For example, there may be an accuracy drop when switching to multi-modal from unimodal models. We will study how to quantify and bound such performance variation by analyzing the information inconsistency between unimodal and multi-modal networks. Moreover, we will also study how to leverage the multi-modal network to improve the performance with only partial data modalities. To this end, we will draw on cross-modal knowledge transfer [54] or data imputation of missing modalities [52].

Dynamic sensor selection. The system efficiency of Harmony can be further improved by exploiting different contributions or strengths of sensor modalities. For example, in federated fusion learning, if the discrepancy of one sensor's unimodal encoder is extremely small for a long period, this sensor does not actually contribute much to the fusion performance on the data of the node's data. We can then turn off the sensor to save computing resources without significantly affecting the model inference accuracy.

10 CONCLUSION

This paper proposes *Harmony*, a new system for heterogeneous multi-modal federated learning. Harmony disentangles multi-modal training in a novel two-stage framework, namely *modality-wise FL* with dynamic resource optimization and *federated fusion learning* by exploiting the modality biases. Extensive experiments on a real-world multi-modal sensor testbed and public datasets show that Harmony significantly outperforms state-of-the-art baselines under various system dynamics.

ACKNOWLEDGMENTS

This work is supported by the Research Grants Council (RGC) of Hong Kong, China, under GRF Grants No. 14203420, the Alzheimer's Drug Discovery Foundation, under Grant RDADB-201906-2019049, the National Natural Science Foundation of China (Project 62271434), Shenzhen Science and Technology Program (Project JCYJ20210324120011032), Guangdong Basic and Applied Basic Research Foundation (Project 2021B1515120008), Shenzhen Key Lab of Crowd Intelligence Empowered Low-Carbon Energy Network (No. ZDSYS20220606100601002), and the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

APPENDIX

The research artifact accompanying this paper is available via https://doi.org/10.5281/zenodo.7927450.

REFERENCES

[1] 2021. Hardware Difference of Tesla Autopilot AP1 vs AP2 vs AP3. https://www.autopilotreview.com/tesla-autopilot-v1-v2-v3-and-beyond-differences/.

[2] 2022. ALZHEIMER'S DIGITAL BIOMARKERS. https://www.alzdiscovery.org/research-and-grants/funding-opportunities/diagnostics-accelerator-digital-biomarkers-program.

[3] 2022. Baidu Apollo. https://www.apollo.auto/.

[4] 2022. NVDIA Xavier NX. https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-xavier-nx/.

[5] 2022. Systemd. https://en.wikipedia.org/wiki/Systemd.

[6] 2022. Waymo Driver. https://waymo.com/waymo-driver/?ncr.

[7] 2023. The CART Home: Collaborative Aging Research using Technology. https://www.ohsu.edu/collaborative-aging-research-using-technology/cart-home.

[8] Ane Alberdi, Alyssa Weakley, Maureen Schmitter-Edgecombe, Diane J Cook, Asier Aztiria, Adrian Basarab, and Maitane Barrenechea. 2018. Smart home-based prediction of multidomain symptoms related to Alzheimer's disease. *IEEE journal of biomedical and health informatics* 22, 6 (2018), 1720–1731.

[9] Manuela Altieri, Federica Garramone, and Gabriella Santangelo. 2021. Functional autonomy in dementia of the Alzheimer's type, mild cognitive impairment, and healthy aging: a meta-analysis. *Neurological Sciences* 42 (2021), 1773–1783.

[10] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. 2019. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818* (2019).

[11] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* 41, 2 (2018), 423–443.

[12] Abdelkareem Bedri, Diana Li, Rushil Khurana, Kunal Bhuwalka, and Mayank Goel. 2020. Fitbyte: Automatic diet monitoring in unconstrained situations using multimodal sensing on eyeglasses. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[13] Chongguang Bi, Guoliang Xing, Tian Hao, Jina Huh, Wei Peng, and Mengyan Ma. 2017. Familylog: A mobile system for monitoring family mealtime activities. In *2017 ieee international conference on pervasive computing and communications (percom)*. IEEE, 21–30.

[14] Joaquim Cerejeira, Luísa Lagarto, and Elizabeta Blagoja Mukaetova-Ladinska. 2012. Behavioral and psychological symptoms of dementia. *Frontiers in neurology* 3 (2012), 73.

[15] Jiayi Chen and Aidong Zhang. 2022. FedMSplit: Correlation-Adaptive Federated Multi-Task Learning across Multimodal Split Networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 87–96.

[16] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, et al. 2019. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2145–2155.

[17] Sijia Chen and Baochun Li. 2022. Towards Optimal Multi-Modal Federated Learning on Non-IID Data with Hierarchical Gradient Blending. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 1469–1478.

[18] Chris Ding and Xiaofeng He. 2004. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*. 29.

[19] Petros Drineas, Alan Frieze, Ravi Kannan, Santosh Vempala, and Vishwanathan Vinay. 2004. Clustering large graphs via the singular value decomposition. *Machine learning* 56 (2004), 9–33.

[20] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. 2020. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948* (2020).

[21] Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. 2020. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems* 33 (2020), 3197–3208.

[22] Guoliang Xing. 2022. Machine Learning Technologies for Advancing Digital Biomarkers for Alzheimer's Disease, Alzheimer's Drug Discovery Foundation. https://www.alzdiscovery.org/research-and-grants/portfolio-details/21130887.

[23] Harish Haresamudram, Irfan Essa, and Thomas Plötz. 2021. Contrastive predictive coding for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–26.

[24] Yuze He, Li Ma, Zhehao Jiang, Yi Tang, and Guoliang Xing. 2021. VI-eye: semantic-based 3D point cloud registration for infrastructure-assisted autonomous driving. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 573–586.

[25] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527* (2016).

[26] Lampros C Kourtis, Oliver B Regele, Justin M Wright, and Graham B Jones. 2019. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *NPJ digital medicine* 2, 1 (2019), 1–9.

[27] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau. 2019. Federated learning for keyword spotting. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6341–6345.

[28] Ang Li, Jingwei Sun, Pengcheng Li, Yu Pu, Hai Li, and Yiran Chen. 2021. Hermes: an efficient federated learning framework for heterogeneous mobile clients. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 420–437.

[29] Chenning Li, Xiao Zeng, Mi Zhang, and Zhichao Cao. 2022. PyramidFL: A fine-grained client selection framework for efficient federated learning. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 158–171.

[30] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems* 2 (2020), 429–450.

[31] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 97–110.

[32] Xiulong Liu, Dongdong Liu, Jiuwu Zhang, Tao Gu, and Keqiu Li. 2021. RFID and camera fusion for recognition of human-object interactions. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 296–308.

[33] Gill Livingston, Andrew Sommerlad, Vasiliki Orgeta, Sergi G Costafreda, Jonathan Huntley, David Ames, Clive Ballard, Sube Banerjee, Alistair Burns, Jiska Cohen-Mansfield, et al. 2017. Dementia prevention, intervention, and care. *The lancet* 390, 10113 (2017), 2673–2734.

[34] Chris Xiaoxuan Lu, Muhamad Risqi U Saputra, Peijun Zhao, Yasin Almalioglu, Pedro PB de Gusmao, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. 2020. milliEgo: single-chip mmWave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 109–122.

[35] Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. 2019. AttnSense: Multi-level Attention Mechanism For Multimodal Human Activity Recognition.. In *IJCAI*. 3109–3115.

[36] Gad A Marshall, Lynn A Fairbanks, Sibel Tekin, Harry V Vinters, and Jeffrey L Cummings. 2006. Neuropathologic correlates of activities of daily living in Alzheimer disease. *Alzheimer Disease & Associated Disorders* 20, 1 (2006), 56–59.

[37] Marie Mc Carthy and P Schueler. 2019. can digital technology advance the development of treatments for Alzheimer's disease? *The Journal of Prevention of Alzheimer's Disease* 6, 4 (2019), 217–220.

[38] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. 2016. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629* (2016).

[39] Hrushikesh Mhaskar, Qianli Liao, and Tomaso Poggio. 2017. When and why are deep networks better than shallow ones?. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[40] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. 2017. CNN-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 158–165.

[41] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, René Vidal, and Ruzena Bajcsy. 2013. Berkeley mhad: A comprehensive multimodal human action database. In *2013 IEEE workshop on applications of computer vision (WACV)*. IEEE, 53–60.

[42] 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.

[43] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Jianwei Huang, and Guoliang Xing. 2021. Clusterfl: a similarity-aware federated learning system for human activity recognition. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 54–66.

[44] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Guoliang Xing, and Jianwei Huang. 2022. ClusterFL: A Clustering-based Federated Learning System for Human Activity Recognition. *ACM Transactions on Sensor Networks* 19, 1 (2022), 1–32.

[45] Ishwari Singh Rajput and Deepa Gupta. 2012. A priority based round robin CPU scheduling algorithm for real time systems. *International Journal of Innovations in Engineering and Technology* 1, 3 (2012), 1–11.

[46] Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine* 34, 6 (2017), 96–108.

[47] 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.

[48] Eric Salmon*, Solange Lespagnard*, Patricia Marique, F Peeters, Karl Herholz, Daniela Perani, Vjera Holthoff, Elke Kalbe, D Anchisi, Stéphane Adam, et al. 2005. Cerebral metabolic correlates of four dementia scales in Alzheimer's disease.

Journal of neurology 252 (2005), 283–290.

[49] Shuyao Shi, Jiahe Cui, Zhehao Jiang, Zhenyu Yan, Guoliang Xing, Jianwei Niu, and Zhenchao Ouyang. 2022. VIPS: real-time perception fusion for infrastructure-assisted autonomous driving. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 133–146.

[50] Jaemin Shin, Yuanchun Li, Yunxin Liu, and Sung-Ju Lee. 2022. FedBalancer: Data and Pace Control for Efficient Federated Learning on Heterogeneous Clients. In *International Conference on Mobile Systems, Applications and Services (MobiSys)*. ACM, 436–449.

[51] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. 2017. Federated multi-task learning. In *Advances in Neural Information Processing Systems*. 4424–4434.

[52] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1405–1414.

[53] Linlin Tu, Xiaomin Ouyang, Jiayu Zhou, Yuze He, and Guoliang Xing. 2021. Feddl: Federated learning via dynamic layer sharing for human activity recognition. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 15–28.

[54] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. 2020. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1828–1838.

[55] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12695–12705.

[56] Fang-Jing Wu and Gürkan Solmaz. 2018. Crowdestimator: Approximating crowd sizes with multi-modal data for internet-of-things services. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 337–349.

[57] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*. PMLR, 24043–24055.

[58] Zhiyuan Xie, Xiaomin Ouyang, Xiaoming Liu, and Guoliang Xing. 2021. UltraDepth: Exposing high-resolution texture from depth cameras. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 302–315.

[59] Zhiyuan Xie, Xiaomin Ouyang, Li Pan, Wenrui Lu, Xiaoming Liu, and Guoliang Xing. 2022. HiToF: a ToF camera system for capturing high-resolution textures. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 764–765.

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

[61] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. LIMU-BERT: Unleashing the Potential of Unlabeled Data for IMU Sensing Applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 220–233.

[62] 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.

[63] Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. 2020. Salvaging federated learning by local adaptation. (2020). arXiv:2002.04758

[64] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* (2014).

[65] Hanbin Zhang, Gabriel Guo, Chen Song, Chenhan Xu, Kevin Cheung, Jasleen Alexis, Huining Li, Dongmei Li, Kun Wang, and Wenyao Xu. 2020. PDLens: smartphone knows drug effectiveness among Parkinson's via daily-life activity fusion. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[66] Mi Zhang and Alexander A Sawchuk. 2012. USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 1036–1043.

[67] 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.