Communication-Efficient Multimodal Federated Learning: Joint Modality and Client Selection

Liangqi Yuan, *Graduate Student Member, IEEE*, Dong-Jun Han, *Member, IEEE*, Su Wang, Devesh Upadhyay, *Senior Member, IEEE*, and Christopher G. Brinton, *Senior Member, IEEE*

***Abstract*—Multimodal federated learning (MFL) aims to enrich model training in FL settings where clients are collecting measurements across multiple modalities. However, key challenges to MFL remain unaddressed, particularly in heterogeneous network settings where: (i) the set of modalities collected by each client is diverse, and (ii) communication limitations prevent clients from uploading all their locally trained modality encoders to the server. In this paper, we propose Multimodal Federated learning with joint Modality and Client selection (MFedMC), a communication-efficient MFL framework that tackles these challenges through a decoupled architecture and selective uploading. Unlike traditional holistic fusion approaches, MFedMC separates modality encoders and fusion modules: modality encoders are aggregated at the server for generalization across diverse client distributions, while fusion modules remain local to each client for personalized adaptation to individual modality configurations and data characteristics. Building on this decoupled design, our joint selection algorithm incorporates two main components: (a) A modality selection methodology for each client, which weighs (i) the impact of the modality, gauged by Shapley value analysis, (ii) the modality encoder size as a gauge of communication overhead, and (iii) the frequency of modality encoder updates, denoted recency, to enhance generalizability. (b) A client selection strategy for the server based on the local loss of modality encoders at each client. Experiments on five real-world datasets demonstrate that MFedMC achieves comparable accuracy to several baselines while reducing communication overhead by over 20×.**

***Index Terms*—Multimodal federated learning, data fusion, Internet of Things, edge computing, communication efficiency.**

I. Introduction

FEDERATED learning (FL) is a distributed machine learning (ML) approach in which users collaboratively train ML models through sharing model parameters rather than raw measurements [2], [3], [4], [5]. The FL approach establishes a federation of learners, each updating their model based on local data. Subsequently, these locally learned parameters are uploaded to a central server or shared with other clients. The local model within each client is then updated by integrating an aggregation of these parameters, ensuring that each model benefits from the collective learning experience. As Internet of Things (IoT) devices, such as smartphones, robots, and autonomous aerial vehicles (AAVs), are increasingly equipped with multimodal sensors, there has been growing interest in multimodal federated learning (MFL) frameworks [6], [7], [8], [9], [10]. For example, in federated settings with connected and automated vehicles (CAVs), diverse sensors such as cameras, LiDAR, and radar provide complementary multimodal measurements that enable robust control decisions across varying weather conditions and fields of view [11], [12], [13].

Multimodal fusion combines information from two or more modalities to enhance ML model performance and robustness through cross-modality interactions [14], [15], [16]. Fusion methods are generally categorized as data, feature, and decision, depending on the network layer where the modalities are merged. Data-level fusion concatenates at the input level (e.g., combining RGB and depth images into RGB-D). Feature fusion merges features from intermediate network layers, enabling nuanced interactions but increasing architectural complexity. Decision fusion aggregates outputs from separate models, offering flexibility but potentially losing fine-grained information during integration. This trade-off between information retention and fusion sophistication is a key consideration in the design of multimodal learning systems.

A. Motivation

In the context of the IoT, where devices measure a diverse set of modalities, most existing MFL frameworks call for massive parallel processing of extensive sensor data streams [17], [18]. They utilize multimodal fusion to boost model performance, especially in scenarios where heterogeneous clients lack one or more modalities. However, IoT devices often possess communication constraints due to bandwidth limitations, location, and varying capabilities. This has also been a key bottleneck in conventional (single modality) FL [19], [20], [21], [22]. Thus, there is still a need to devise strategies to reduce communication overhead and improve learning efficiency within the MFL

Received 1 July 2025; revised 22 December 2025; accepted 7 March 2026. Date of publication 13 March 2026; date of current version 6 July 2026. This work was supported by the Office of Naval Research (ONR) under Grant N00014-23-C-1016, in part by the National Science Foundation (NSF) under Grant CPS-2313109, and in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-24-1-0083. An earlier version of this paper was presented at the part of 2024 IEEE International Conference on Communications (ICC) [DOI: 10.1109/ICC51166.2024.10622194]. Recommended for acceptance by X. Chen. *(Corresponding author: Liangqi Yuan.)*

Liangqi Yuan and Christopher G. Brinton are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907 USA (e-mail: liangqiy@purdue.edu; cgb@purdue.edu).

Dong-Jun Han is with the Department of Computer Science and Engineering, Yonsei University, Seoul 03722, South Korea (e-mail: djh@yonsei.ac.kr).

Su Wang is with the School of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08540 USA (e-mail: hw5731@princeton.edu).

Devesh Upadhyay is with Saab Inc., East Syracuse, NY 13057 USA (e-mail: deveshu@gmail.com).

A demo video and our code are available at https://liangqiy.com/mfedmc/.

Digital Object Identifier 10.1109/TMC.2026.3673697

paradigm. We summarize the existing gaps and challenges in the current MFL frameworks as follows.

i) Client and Modality Heterogeneity extends beyond statistical heterogeneity (non-IID data distributions) to encompass individual, group, and system differences [23], [24], [25]. Individual heterogeneity manifests in feature variations (e.g., walking frequencies), group heterogeneity arises from physiological differences (e.g., left-hander vs. right-hander), and system heterogeneity stems from external factors (e.g., device age). Critically, in MFL, modality heterogeneity is prevalent where some clients may lack certain modalities entirely (e.g., some CAVs are equipped with LiDAR while others rely solely on vision-based systems), posing unique challenges in MFL scenarios.

ii) Communication Efficiency is influenced not only by the balance between performance and communication overhead, but also by the varying data sizes across different modalities and data types, leading to differences in modality encoder sizes. Furthermore, the complexity of information patterns inherent in different data modalities affects the ease of learning. For instance, compared to high-resolution image data, simpler data modalities like time-series radar data, despite potentially lower accuracy, may require only simple ML models for effective training and recognition due to their smaller data size and straightforward pattern recognition.

iii) Impact of Modality and Client in MFL are critical yet under explored areas, particularly in the context of communication costs and client participation. It remains an open question as to which modality or combination of modalities should be prioritized in the predictions, considering their information richness. Moreover, the extent of client participation, influenced by factors such as data availability and quality, also plays a vital role in determining the effectiveness of different modalities within the MFL framework.

Motivated by these challenges, we aim to answer the following key questions:

> In resource-constrained and heterogeneous MFL settings, (i) how should we design the learning architecture to handle clients with heterogeneous modality sets while enabling both generalization and personalization, (ii) how should each client evaluate and select the most valuable modalities for uploading to balance performance and communication overhead, and (iii) how should the server determine which clients, each with different modality encoder combinations, to aggregate from to further optimize this trade-off?

B. Overview and Contribution

We propose Multiodal Federated learning with joint Modality and Client selection (MFedMC), an MFL methodology tailored for clients with heterogeneously absent modalities. An overview of our proposed method and comparison with traditional MFL are given in Fig. 1. We propose to decouple modality encoders and fusion modules, where *global modality encoders* generate predictions that serve as input to the individual *local fusion module* in each client. This modular design renders our framework

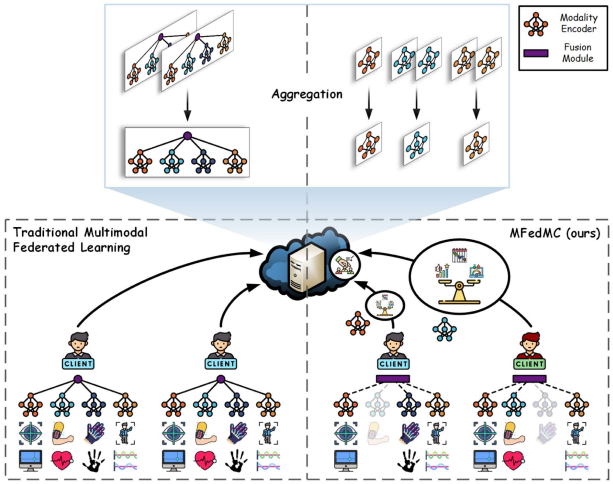


Fig. 1. Comparison between traditional multimodal federated learning vs. the proposed MFedMC.

adaptable and beneficial across various applications that involve multiple modalities, thus facilitating the collaboration between clients. For example, CAVs utilize a combination of cameras, LiDAR, and radar, each providing complementary information essential for autonomous driving. Our framework supports modular modality encoders on vehicles equipped with these different sensors, enabling effective MFL across different vehicles. While modality encoders are uploaded to the server for aggregation to enhance generalization, each client retains the fusion module locally to improve personalization and confidence. Retaining the fusion module local to the client minimizes the risk of leakage of client sensitive information to the server either directly or indirectly via server side inference.

Building on this foundation, we introduce the concept of *joint modality and client selection* to reduce communication overhead. First, *selective modality communication* acknowledges that clients may not always have the capacity or necessity to upload all modality encoders (e.g., resource-constrained IoT applications). This approach is based on factors such as modality performance or its impact on the final decision, communication overhead, and recency. To evaluate the impact of each modality, we propose employing Shapley values [26], [27], [28], [29], [30], measured on the fusion module, to quantify their respective performances. Communication overheads are determined by quantifying the size of modality encoders, while recency is gauged by tracking the upload history of modalities. Complementing this, *selective client uploading* is proposed to further reduce communication overheads and address the effects of client heterogeneity on the global model. This involves ranking each client based on the local loss of modality encoder, thereby optimizing both the efficiency and effectiveness of our FL framework.

In this paper, we present the following contributions:

- *Multimodal Federated Learning with Decoupled Encoders and Fusion (Sec. III-A):* We propose a framework that decouples traditional holistic fusion approaches into separate modality encoders and fusion modules. Modality encoders

are uploaded to a server to enhance generalizability, while individual fusion modules are retained locally to promote personalization and strengthen confidence. This decoupled approach enables modular functionality of modalities and naturally accommodates client heterogeneity and missing modality scenarios.

- *Communication-Efficient Joint Modality and Client Selection (Sec. III-B, Sec. III-C):* We propose employing a Shapley component to quantify the impact of modality encoders, along with assessing their sizes to estimate communication overhead. Additionally, a recency term is introduced to track the frequency of uploads, facilitating selective modality communication. Furthermore, we implement selective client uploading, which leverages the local loss of modality encoder to quantify client heterogeneity. This joint modality and client selection strategy significantly reduces communication costs while maintaining model performance, thereby enhancing learning efficiency.
- *Five Real-World Experiments (Sec. IV-C):* Our proposed MFedMC is evaluated for performance and communication overhead against four baseline methods across five heterogeneous multimodal datasets, including wearable sensors, healthcare, language, and satellite datasets. Experiment results highlight the superior performance of MFedMC, achieving comparable accuracy while incurring less than 25% of the communication overhead compared to baseline methods.
- *Analytics on Modality Impact (Sec. IV-D):* We offer analyses on modality impact within the FL process. Utilizing Shapley values, we illustrate the interplay among modalities, revealing the dynamic impact of each modality on the final decision-making process within scenarios that take into account communication overhead and recency.

This paper is an extension of our previous paper [1]. Building upon the foundation laid by [1], this paper introduces the following key contributions: (i) For modality selection, we incorporate a new metric, recency, to prevent overemphasis on certain modalities and maintain generalization. (ii) We introduce a client selection strategy tailored to MFL that synergistically optimizes communication overhead in conjunction with modality selection. (iii) Beyond the previously experimented ActionSense dataset, we have added four new datasets, covering domains such as wearable sensors, healthcare, language, and satellite imagery, providing a holistic demonstration of the framework's applicability. (iv) We extend our experimental suite with additional baselines and comprehensive ablation studies across diverse federated settings, including class heterogeneity, modality heterogeneity, long-tailed distributions, and client dynamics.

C. Organization

The rest of this paper is organized as follows. Section II reviews related works, and Section III details our MFedMC algorithm. We present our experiments and their corresponding results in Section IV. Finally, we summarize our conclusions and future work in Section V.

II. Related Works

A. Multimodal Federated Learning

MFL has gained increasing attention due to the inherently multimodal nature of real-world data and the complementary benefits offered by leveraging multiple modalities. Recently, a variety of algorithms have been proposed to improve the performance of MFL frameworks based on different fusion techniques, including data-level [31], feature-level [32], and decision-level [33]. However, due to client heterogeneity, different clients may possess various modalities, which is also known as modality heterogeneity challenge [18], [34], [35], [36]. Most current MFL implementations are confined to end-to-end model architectures, resulting in either (i) inability to aggregate on the server because of different model architectures caused by varying modalities or (ii) significant performance degradation when using zero padding or random padding to compensate for missing modality data [33], [37]. Several related works aim to address these challenges, such as utilizing pre-trained autoencoders to generate missing modalities [38], relying on a complete public dataset at the server [39], cross-modality reconstruction to recover missing modalities [40], and others.

However, our approach circumvents the performance degradation associated with zero padding or random padding techniques by decoupling the learning process to focus exclusively on available modalities at each client. At the server side, we aggregate modality-specific encoders rather than complete models. The fusion modules remain localized due to structural variations across clients and serve as personalization mechanisms, as global encoders may not align with local data distributions. For example, a left-handed user's client would likely experience suboptimal performance with a global model predominantly trained on right-handed users' data, thus the fusion module functions to appropriately attenuate the weights of potentially misaligned modality encoders. Furthermore, our modality selection strategy leverages the decoupled fusion module to quantify the impact contribution of each modality at the client level.

B. Selection Strategies in Federated Learning

Multi-objective selection in network scenarios is not uncommon, with numerous related works selecting devices for resource optimization based on different trade-offs [41]. In FL, due to the extensive heterogeneity among clients, including information richness, information asymmetry, and security and privacy concerns, related works have proposed static or dynamic client selection functions for uploading and aggregation [42], [43], [44], [45], such as network heterogeneity [46], client dataset similarity [47], client training loss [48], and other schemes. In the MFL scenario, beyond the aforementioned client heterogeneity, a new heterogeneity challenge emerges: modality heterogeneity among clients. To date, within the MFL context, a preliminary test in FLASH [49] involves clients randomly selecting and uploading one of three modality encoders or a fusion module for aggregation.

However, no research has yet explored modality selection in the MFL context, primarily because modality encoders

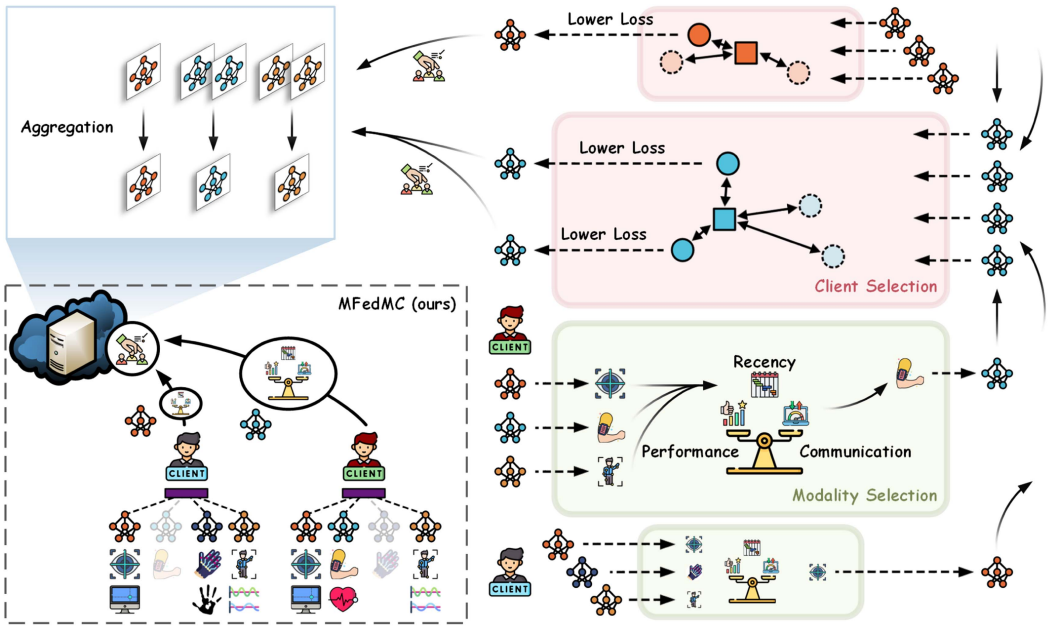


Fig. 2. System diagram of the proposed MFedMC illustrating the process of **Modality Selection** and **Client Selection** as detailed in Algorithm 1.

differ in training difficulty and impact during fusion. Moreover, modality encoders evidently possess different parameter quantities, further resulting in varying communication overheads. Our approach addresses the multi-objective optimization problem of performance-communication trade-offs, dynamically selecting modalities during the MFL process. We also highlight the dynamic selection process in MFL, where selected modalities should differ in each communication round to maximize modality complementarity. Our method adaptively selects different modalities at various stages of the dynamic MFL process, choosing more easily trainable modalities in early stages and more difficult but information-rich modalities in later stages. This not only ensures the convergence speed of the MFedMC framework but also guarantees its overall superior accuracy.

III. Formulation and Methodology

A. Proposed MFedMC Framework

System Model: Consider an MFL system with K clients. Each client k possesses a local dataset $\mathbb{D}^k$ with corresponding labels Y^k, comprising multimodal data represented as

$$\mathbb{D}^k = \{\mathcal{D}_1^k, \mathcal{D}_2^k, \ldots, \mathcal{D}_{M_k}^k\}, \tag{1}$$

where $\mathcal{D}_m^k$ denotes the dataset for modality m (e.g., text, images, LiDAR), and M_k represents the number of available modalities at client k. Note that M_k may vary across clients due to modality heterogeneity, where some clients may lack certain sensing capabilities.

Decoupled Architecture: To accommodate class and modality heterogeneities, we decouple the multimodal learning process into two components: global modality encoders and local fusion modules. Each client k maintains a set of modality encoders

$$\Theta^k = \{\theta_1^k, \theta_2^k, \ldots, \theta_{M_k}^k\}, \tag{2}$$

where each encoder θ_m^k extracts discriminative features from its corresponding modality dataset $\mathcal{D}_m^k$ and produces predictions. Specifically, for each modality m, the encoder generates a predicted label

$$\hat{y}_m^k = \theta_m^k(\mathcal{D}_m^k), \tag{3}$$

$$\widehat{\mathbb{Y}}^k = \{\hat{y}_1^k, \hat{y}_2^k, \ldots, \hat{y}_{M_k}^k\}, \tag{4}$$

where $\widehat{\mathbb{Y}}^k$ denotes the collection of predictions from all modality encoders at client k. A local fusion module ω^k then aggregates these predictions to produce the final output:

$$\begin{aligned}\widehat{Y}^k &= \omega^k\left(\theta_1^k(\mathcal{D}_1^k), \theta_2^k(\mathcal{D}_2^k), \ldots, \theta_{M_k}^k(\mathcal{D}_{M_k}^k)\right) \\ &= \omega^k\left(\hat{y}_1^k, \hat{y}_2^k, \ldots, \hat{y}_{M_k}^k\right) = \omega^k\left(\widehat{\mathbb{Y}}^k\right).\end{aligned} \tag{5}$$

where $\widehat{Y}^k$ represents the final prediction for client k. This decoupled design offers several advantages: (i) modality encoders θ_m can be collaboratively trained across clients and shared via the server to enhance generalizability, (ii) fusion modules ω^k remain local to each client, enabling personalized adaptation to individual heterogeneities such as modality availability, user characteristics, and noise levels, and (iii) the modular structure naturally accommodates missing modality scenarios without requiring architectural modifications.

Learning Objectives: At each client k, all modality encoders are trained in parallel, with each encoder independently minimizing the discrepancy between its predicted labels and the true labels using a loss function f:

$$\min_{\theta_m^k} f(Y^k, \hat{y}_m^k), \quad \forall m \in \{1, 2, \ldots, M_k\}. \tag{6}$$

Subsequently, the fusion module aims to minimize the discrepancy between the true labels and the fused predictions across all modalities:

$$\min_{\omega^k} f(Y^k, \widehat{Y}^k). \tag{7}$$

Algorithm 1: MFedMC: Multimodal Federated Learning with Joint Modality Selection and Client Selection.

Input: Communication rounds (T), client datasets ($\{\mathbb{D}^k\}_{k=1}^K$), client label sets ($\{Y^k\}_{k=1}^K$), local training epoch (E), initial models ($\theta_{m,0}^k$ and ω_0^k), loss function (f), learning rate (η), modality encoder upload count (γ), modality selection weights (α_s, α_c, and α_r), client selection ratio (δ)
Output: Generalized global modality encoders (θ_m) and personalized local fusion modules for each client (ω^k)

 # Global Iteration
for $t = 0$ **to** $T-1$ **do**

 # Local Learning
for each client k **in parallel do**
  for each data modality $m \in \mathcal{M}^k$ **in parallel do**
    Train modality encoder: $\theta_m^{k,t+1} \leftarrow \arg\min_{\theta_m^{k,t}} f(Y^k, \hat{y}_m^{k,t})$ for E epochs
  end for
  (Stage #1) Freeze encoders $\{\theta_m^{k,t+1}\}_{m\in\mathcal{M}^k}$ and train fusion module $\omega^{k,t+1}$ for E epochs
end for

 # Modality Selection
Compute the Shapley values for each modality $\Phi^{k,t}$ ▷ (8), (9)
Compute the modality encoder size $\bar{\Theta}^k$ ▷ (10)
Compute the recency $\mathcal{T}_m^k$ ▷ (11)
Clients select the modality encoders with top-γ priority for uploading and update recency ▷ (14), (15), (16)

 # Client Selection & Server Aggregation
Server selects top-δ clients with lowest modality encoder losses ▷ (18), (19), (20)
for each data modality $m \in \mathcal{M}$ **do**
  Aggregate uploaded modality encoder θ_m^{t+1} ▷ (21)
end for

 # Local Deploying
for each client k **in parallel do**
  for each data modality $m \in \mathcal{M}^k$ **in parallel do**
    Download and deploy global encoder θ_m^{t+1}
  end for
  (Stage #2) Freeze encoders $\{\theta_m^{t+1}\}_{m\in\mathcal{M}^k}$ and fine-tune fusion module $\omega^{k,t+1}$ for E epochs
end for

end for
return Global encoders $\{\theta_m^T\}_{m\in\mathcal{M}}$ and local fusion modules $\{\omega^{k,T}\}_{k=1}^K$

Joint Selection Overview: The MFedMC algorithm, illustrated in Fig. 2 and detailed in Algorithm 1, achieves communication-efficient collaborative learning through joint client and modality encoder selection. During each communication round, rather than requiring all clients to upload all modality encoder parameters, MFedMC selects a subset of clients and determines which modality encoders each selected client should upload. This selective aggregation substantially reduces communication overhead while maintaining model performance. The server aggregates the received encoder parameters and broadcasts the updated global encoders back to clients, which then perform local training on both encoders and fusion modules. Within a fixed communication budget, MFedMC maximizes model accuracy by intelligently determining which clients and modality encoders contribute most effectively to the global model in each round. In the following sections, we provide detailed descriptions of the joint selection mechanism and training procedure.

B. Modality Selection

Due to the resource constraints on edge devices serving as clients, they may not possess adequate storage capacity to house extensive multimodal data, computational capability to learn on multimodal datasets, or communication bandwidth to upload several models to the server. Thus, we introduce three metrics to assist clients to determine whether to upload their models to the server:

- *Shapley value (φ)* represents the impact of a modality encoder on the final prediction, where a higher value of φ corresponds to a higher priority ↑.
- *Modality encoder size ($|\theta|$)* pertains to the communication overhead, where a lower value of $|\theta|$ corresponds to a higher priority ↓.
- *Recency ($\mathcal{T}$)* denotes the freshness or recentness of a client's modality encoder upload, where a higher value indicates that the modality has not been updated recently, thus it is given higher priority ↑.

Shapley Value (Impact of Modality): During each communication round, clients evaluate the impact of the modality encoders Θ^k on the outcomes utilizing interpretability techniques and choose to upload only a selected subset of encoders. We consider using the Shapley value as an assessment to evaluate the relationship between input $\widehat{\mathbb{Y}}^k$ and output $\widehat{Y}^k$ of the fusion module ω^k:

$$\varphi_m^k = \sum_{\mathcal{Y}\subseteq\widehat{\mathbb{Y}}^k\setminus\{m\}} \frac{|\mathcal{Y}|!(|\widehat{\mathbb{Y}}^k| - |\mathcal{Y}| - 1)!}{|\widehat{\mathbb{Y}}^k|!} \times \left(\omega^k(\mathcal{Y}\cup\{m\}) - \omega^k(\mathcal{Y})\right), \tag{8}$$

where φ_m^k is the Shapley value of input modality m, $\mathcal{Y}$ is a subset of $\widehat{\mathbb{Y}}^k$ excluding modality m, and $\omega^k(\mathcal{Y})$ is the predicted value using only modalities in set $\mathcal{Y}$. For all modalities, we assess the magnitude of each Shapley value by taking its absolute value and construct the following set:

$$\Phi^k = \left\{|\varphi_1^k|, |\varphi_2^k|, \ldots, |\varphi_{M_k}^k|\right\}. \tag{9}$$

Modality Encoder Size (Communication Overhead): Given modality encoders with parameters $\Theta^k = \{\theta_1^k, \theta_2^k, \ldots, \theta_{M_k}^k\}$, the communication overhead for each modality encoder is directly proportional to the model size given by

$$\bar{\Theta}^k = \left\{|\theta_1^k|, |\theta_2^k|, \ldots, |\theta_{M_k}^k|\right\}. \tag{10}$$

Recency: Our primary aim with the recency metric is to encourage clients to update a specific modality encoder to the server. This is to ensure that the MFedMC framework doesn't overly prioritize data modalities that are more easily obtainable, possess simpler data structures, or have conspicuous features, thus sidelining the diversity of other data modalities. To encapsulate the concept of recency in our model, we denote $\mathcal{T}_m^k$ as:

$$\mathcal{T}_m^k = t - t_m^k - 1, \tag{11}$$

where $t = 1, 2, \ldots, T$ represents the current communication round, and t_m^k indicates the communication round during which the modality encoder from modality m of client k was last uploaded.

Priority (Composite Score): Considering the impact of the modality encoder, as quantified by the Shapley value, the communication overhead as characterized by the modality encoder size, and the timeliness of model updates captured by the recency, we propose priority P as a composite score. To derive the priority, we proceed with individual normalization for each criterion:

$$\begin{cases} \tilde{\varphi}_m^k = \frac{\varphi_m^k - \min(\Phi^k)}{\max(\Phi^k) - \min(\Phi^k)}, \\ |\tilde{\theta}_m^k| = \frac{\theta_m^k - \min(\bar{\Theta}^k)}{\max(\bar{\Theta}^k) - \min(\bar{\Theta}^k)}, \text{ for } m = 1, 2, \ldots, M_k, \\ \tilde{\mathcal{T}}_m^k = \frac{\mathcal{T}_m^k}{t}, \end{cases} \tag{12}$$

where $\tilde{\varphi}_m^k$ represents the normalized Shapley Value, $|\tilde{\theta}_m^k|$ denotes the normalized communication overhead, and $\tilde{\mathcal{T}}_m^k$ indicates the normalized recency. With a focus on identifying modality encoders for server communication, we devise the priority P_m^k for each modality and the corresponding set $\mathcal{P}^k$ to determine whether modality encoders should be sent to the server. They are formulated as:

$$\begin{aligned} P_m^k &= \alpha_s \cdot \tilde{\varphi}_m^k + \alpha_c \cdot (1 - |\tilde{\theta}_m^k|) + \alpha_r \cdot \tilde{\mathcal{T}}_m^k, \\ \mathcal{P}^k &= \left\{P_1^k, P_2^k, \ldots, P_{M_k}^k\right\}, \end{aligned} \tag{13}$$

where α_s, α_c, and α_r are the predetermined metric weights, satisfying $\alpha_s + \alpha_c + \alpha_r = 1$. Naturally, a modality with the maximal priority is considered optimal.

Modality Selection: To streamline our decision-making, we focus on modalities with the highest γ priorities:

$$\mathcal{P}_\gamma^k = \text{top}_\gamma \max(\mathcal{P}^k) \tag{14}$$

Hence, the selected modality set of client k becomes:

$$\mathcal{M}_\gamma^k = \left\{m \,:\, P_m^k \in \mathcal{P}_\gamma^k, \text{ for } m = 1, 2, \ldots, M_k\right\}. \tag{15}$$

With $\mathcal{M}_\gamma^k$ determined, the set of modality encoders ready for communication to the server by the client k can be defined as:

$$\Theta_\gamma^k = \left\{\theta_m^k : \theta_m^k \in \Theta^k \text{ and } m \in \mathcal{M}_\gamma^k\right\}, \tag{16}$$

where the set Θ_γ^k represents the modality encoders corresponding to the top-γ priority from $\mathbb{S}^k$. Each client will upload a data packet with various details to the server for aggregation, including model parameters θ^k, modality information m, the number of samples $|\mathcal{D}_m^k|$, among others. Likewise, upon downloading from the server, this information will also be retrieved. Note that only model θ^k will be uploaded/downloaded to/from the server. The fusion module ω^k varies across clients, determined by the unique deployment scenarios of each client, such as geographical location, operational duration, external interference, etc.

C. Client Selection and Joint Selection

Client Selection: Considering the heterogeneity of clients in practice, the proposed MFedMC is designed to further optimize communication overhead and enhance learning efficiency through a client selection strategy. During the training process of the client models, we consider loss ℓ as a selection criterion, focusing on the modalities chosen to communicate to the server. For each modality m, the set $\mathcal{L}_\gamma$ includes the loss values ℓ_m^k from each client k that has modality m in their selected set $\mathcal{M}_\gamma^k$:

$$\mathcal{L}_\gamma = \left\{\ell_m^k \mid m \in \mathcal{M}_\gamma^k \text{ and } k = 1, 2, \ldots, K\right\}. \tag{17}$$

This collection $\mathcal{L}_\gamma$ allows the server to make informed decisions about which clients to select to participate in the learning process, based on the reported loss values across the different modalities. The set of the lowest $\lceil \delta K \rfloor$ loss values can be expressed as:

$$\mathcal{L}_\gamma^\delta = \text{top}_{\lceil \delta K \rfloor} \min(\mathcal{L}_\gamma) \tag{18}$$

where the ratio parameter δ establishes the proportion of the lower loss values to be considered. This strategy utilizes the top-($\lceil \delta K \rfloor$) criterion to pick a subset of clients, concentrating on those with the lower loss values, which are indicative of their potential effect on the learning efficiency. $\lceil \cdot \rfloor$ denotes the rounding operation. Hence, the set of selected clients can be expressed as:

$$\mathcal{K}_\gamma^\delta = \left\{k \,:\, \ell^k \in \mathcal{L}_\gamma^\delta, \text{ for } k = 1, 2, \ldots, K\right\}. \tag{19}$$

Joint Selection: In summary, the joint selection process is completed by first identifying the modality set $\mathcal{M}_\gamma^k$ for each client k, followed by selecting the client set $\mathcal{K}_\gamma^\delta$ based on the loss values. This initial selection prioritizes modalities that are most significant for each client. Subsequently, client selection identifies clients based on the loss. This sequential selection process ensures that we first optimize the selection of modalities before determining which clients will contribute to the learning process. The identification of the client set $\mathcal{K}_\gamma^\delta$ and the final

model set to communicate to the server is then achieved as:

$$\Theta_{\gamma}^{\delta} = \left\{ \theta_m^k : \theta_m^k \in \Theta_{\gamma}^k \text{ and } k \in \mathcal{K}_{\gamma}^{\delta} \right\}, \tag{20}$$

where the set Θ_{γ}^k represents the modality encoders corresponding to the top-γ priority defined in (16). Through joint modality and client selection, the expected reduction in communication overhead is quantified by the ratio $\gamma/\bar{M} \times \delta$, where $\bar{M} \geq 2$ denotes the average number of modalities per client. The factor $0 < \gamma/\bar{M} < 1$ represents the communication overhead reduction achieved through modality selection, while $0 < \delta < 1$ represents the communication overhead reduction achieved through client selection. To illustrate, consider an MFL environment comprising 100 clients, each utilizing an average of 3 modalities. In a conventional MFL paradigm, each training round necessitates 300 modality encoder uploads (100 clients $\times$ 3 modalities). Employing our joint selection strategy with parameters $\gamma = 1$ and $\delta = 0.2$, the system requires merely 20 modality encoder uploads each round, constituting a 93.3% reduction in communication overhead. With this strategic joint selection approach, the observed performance degradation is significantly smaller than the corresponding reduction in communication overhead, thus effectively achieving our objective.

D. Server Aggregation and Local Retention

Server Aggregation: Learning Generalizable Modality Encoders: To achieve robust cross-client generalization, modality encoders are aggregated at the server. Upon receiving the selected encoders from participating clients, the server performs weighted aggregation for each modality m as

$$\theta_m \leftarrow \sum_{\theta_m^k \in \Theta_{\gamma}^{\delta}} \frac{|\mathcal{D}_m^k|}{\sum_{k=1}^{K_m} |\mathcal{D}_m^k|} \theta_m^k, \tag{21}$$

where K_m denotes the number of clients that uploaded the encoder θ_m in the current round, and $|\mathcal{D}_m^k|$ represents the sample count of modality m in client k.

Local Retention: Enabling Personalized Fusion Modules: In stark contrast, fusion modules ω^k remain strictly local and are never transmitted to the server. This deliberate design choice enables personalization: each fusion module specializes in integrating modality predictions according to its unique local context, including modality heterogeneity, user-specific patterns, device characteristics, and individual data distributions. The fusion module undergoes a two-stage training protocol:

- *Stage #1:* After local encoder training, the fusion module is trained with frozen encoders to calculate Shapley values for modality selection.
- *Stage #2:* Upon receiving updated global encoders, each client fine-tunes its fusion module to learn optimal integration strategies tailored to local conditions.

Modality Scalability of Our Decoupled Architecture: Our decoupled architecture naturally accommodates dynamic modality configurations. When a client's modality set changes (e.g., adding or removing sensors), it simply downloads the corresponding global encoders from the server, adjusts the fusion module architecture, and retrains only the lightweight fusion module with frozen encoders, without retraining the entire model. In terms of computational cost, training modality encoders incurs the same overhead as classical multimodal learning, and since encoders are trained in parallel (Algorithm 1), the wall-clock training time is comparable to classical approaches. The additional overhead stems from Shapley value calculation, which grows exponentially with modalities ($\mathcal{O}(M_k 2^{M_k})$). To address this, we employ tree-based fusion modules and interventional feature perturbation [30] to reduce complexity from $\mathcal{O}(N_{\omega^k} L_{\omega^k} M_k 2^{M_k})$ to $\mathcal{O}(N_{\omega^k} L_{\omega^k} H_{\omega^k} |D_k|)$, where N_{ω^k}, L_{ω^k}, H_{ω^k}, and $|D_k|$ denote the number of trees, leaves, maximum depth, and background dataset size. We further subsample background data ($|D_k'| \ll |D_k|$) for efficient estimation, ensuring MFedMC scales effectively as clients incorporate increasing numbers of modalities.

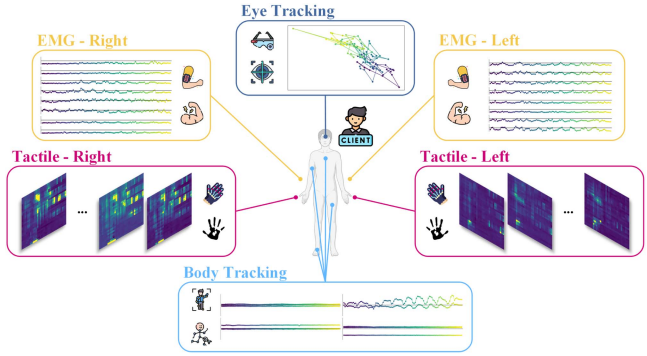


Fig. 3. Data visualization of ActionSense dataset for subject 00 engaged in peeling a potato.

IV. Experiment and Results

A. Dataset

Our experiments are conducted across five diverse multimodal real-world datasets, each incorporating varying numbers of clients, modalities, application contexts, and data types and structures, as depicted in Table I. Specifically, we consider:

i) ActionSense [50] is an extensive multimodal dataset that captures human daily activities. It integrates data from six different types of wearable sensors, documenting human interactions with objects and the environment in a kitchen context. Activities include tasks such as peeling a cucumber, slicing a potato, cleaning a plate with a sponge, and so forth. Fig. 3 illustrates the six modality data from the ActionSense dataset for Subject 00 while peeling a potato.

ii) UCI-HAR [51], similar to ActionSense, employs wearable sensors to monitor people during routine activities such as Walking, Walk Downstairs, Sitting, etc. Compared to ActionSense, UCI-HAR features a larger number of subjects but fewer modalities.

iii) PTB-XL [52], a substantial electrocardiography (ECG) dataset, comprises diverse clinical patient data, featuring 12-lead ECGs from various hospitals. Following established protocols in [53], we treat the Limb Lead ECG and the Precordial Lead ECG as separate modalities. Each data entry includes five labels: Normal, Myocardial Infarction, ST/T Change, Conduction Disturbance, and Hypertrophy.

TABLE I
DESCRIPTION OF DATASETS

Dataset	Client	Task	Modality	Feature
ActionSense [1]	9 Subjects	20 Kitchen Activities	Eye Tracking EMG ($\times 2$) (Left and Right Arm) Tactile ($\times 2$) (Left and Right Hand) Body Tracking	2 $8(\times 2)$ $32 \times 32(\times 2)$ 22×3
UCI-HAR	30 Subjects	6 Daily Activities	Accelerometer Gyroscope	128×3 128×3
PTB-XL	39 Hospitals	5 Diagnosis	Limb Lead ECG [2] Precordial Lead ECG [3]	1000×6 1000×6
MELD	42 Speakers	4 Emotions	Audio Text	Client max length$\times 80$ [4] 100
DFC2023	10 Cities of GF2 Satellite + 17 Cities of SV Satellite	12 Roof Types	SAR Optical	$32 \times 32 \times 1$ $32 \times 32 \times 3$

[1] Subjects 06 through 09 miss both left and right tactile data.
[2] Limb Lead ECG refers to I, II, III, aVL, aVR, aVF leads.
[3] Precordial Lead ECG refers to V1–V6 leads.
[4] Client max length refers to the max length of audio utterances for each client.

iv) MELD [54] is a natural language processing (NLP) dataset derived from dialogues in the TV-series Friends. In each dialogue scene, each speaker is represented by audio and text modalities, reflecting different emotions such as neutral, sadness, joy, and anger.
v) The 2023 IEEE GRSS Data Fusion Contest (DFC23) [55] is a large-scale dataset featuring rooftop satellite imagery of buildings. It contains images from the Gaofen-2 (GF2) and SuperView-1 (SV) satellites, covering rooftops in seventeen cities across six continents. The dataset includes synthetic aperture radar (SAR) and optical images, categorizing twelve types of rooftops.

We test our method against baselines on the aforementioned five datasets, considering two primary evaluation metrics: (i) accuracy under communication constraints and (ii) communication overhead required to achieve target accuracy. We establish target accuracy thresholds of 85% , 60% , 50% , 50% , and 60% for ActionSense, UCI-HAR, PTB-XL, MELD, and DFC23 datasets, respectively. On this foundation, we conduct extensive experiments comparing our method with baselines under the following four client distribution scenarios:

- In the natural distribution setting, we maintain the original client division inherent in the datasets that authentically reflects real-world scenarios, designating entities such as human subjects, speakers, hospitals, and satellites as distinct clients. This setting naturally embodies system-level heterogeneities inherent to individual clients, including variations in noise levels, sensor calibration differences, device age, data collection protocols, and environmental conditions across different deployment sites.
- In the IID setting, we shuffle all data and then redistribute these shuffled samples to all clients. In this scenario, individual, group, and system heterogeneities are significantly diminished due to the random and uniform distribution of samples.
- In the class non-IID setting, we allocate data to clients according to a Dirichlet distribution Dir(β), where β is the concentration parameter controlling the degree of class imbalance. A smaller β value results in more heterogeneous class distributions across clients, with each client predominantly receiving samples from a limited subset of classes.
- In the modality non-IID setting, we randomly remove modalities from clients at certain probabilities. This leads to each client having a different combination of models while missing some modalities.

B. Experiment Setup

Training Setup: To ensure a fair comparison across datasets (i – iv), we initially reshape all data modalities into a two-dimensional format, i.e., time $\times$ features. For all modalities, we employ a consistent long short-term memory (LSTM) network structure, consisting of a single LSTM layer with 128 hidden units, followed by a fully connected layer. The learning rate η for these LSTM models is set to 0.1. For the dataset (v), DFC23, which is composed entirely of images, we use a convolutional neural network (CNN) with one 5×5 convolution layer containing 32 channels, followed by a rectified linear unit (ReLU) activation, a 2×2 max pooling, and a fully connected layer. The learning rate η for these CNN models is set at 0.01. We adopt cross-entropy loss with stochastic gradient descent (SGD) as the optimizer, a batch size of 32, and a local training epoch $E = 5$.

Baselines: At a high-level, we use two types of baselines for comprehensive comparisons: state-of-the-art (SOTA) MFL frameworks, and ablation studies of our method. We include five SOTA MFL frameworks in our analysis: FL-FD [31], MMFed [32], FedMultimodal [36], FLASH [49], and Harmony [56]. Furthermore, to validate the effectiveness of the proposed MFedMC, we conduct comprehensive ablation studies in which the modality selection strategy, client selection strategy, and joint selection strategy are systematically replaced with random selection mechanisms. To ensure a fair systematic comparison within the FL context, we maintain consistent configurations across all methods, including identical base network architectures (LSTM for datasets (i – iv), CNN for dataset (v)), uniform hyperparameters (learning rate, batch size, optimizer, loss function, local training epoch, etc.), among others. We do not incorporate various specialized techniques present in the original baseline papers, such as co-attention mechanisms,

TABLE II
MAIN RESULTS - OVERALL COMPARISON WITH BASELINES: (I) ACCURACY UNDER 5 MB COMMUNICATION CONSTRAINT AND (II) COMMUNICATION OVERHEAD TO ACHIEVE TARGET ACCURACY

	(i) Accuracy (%) ↑					(ii) Communication Overhead (MB) ↓				
	Action Sense	UCI-HAR	PTB-XL	MELD	DFC23	Action Sense	UCI-HAR	PTB-XL	MELD	DFC23
IID Setting										
FL-FD [31]	39.50	35.89	22.28	50.30	62.49	48.71	50.13	N/A	16.89	4.54
MMFed [32]	44.04	21.34	13.31	51.85	61.81	82.34	152.37	N/A	18.09	9.04
FedMultimodal [36]	13.45	17.52	4.07	49.39	60.24	135.08	143.70	N/A	271.32	9.04
FLASH [49]	21.41	22.88	17.44	49.10	61.67	N/A	N/A	N/A	N/A	9.04
Harmony [56]	42.02	21.61	15.84	45.26	57.83	110.24	145.22	N/A	N/A	18.04
Ours w/o Modality Sel	68.73	76.92	86.50	54.89	61.16	5.99	**5.78**	**0.24**	2.51	9.21
Ours w/o Client Sel.	81.79	70.90	85.66	55.58	63.81	N/A	12.96	**0.24**	**0.37**	4.18
Ours w/o Joint Sel.	68.00	71.36	86.90	55.44	**64.36**	15.70	9.11	**0.24**	2.51	**0.84**
MFedMC (Ours) [1]	**92.28** [2]	**78.10**	**87.04**	**55.82**	62.05	**4.27**	7.18	**0.24**	**0.37**	9.20
Natural Distribution										
FL-FD [31]	51.57	35.89	22.28	51.73	62.49	48.71	50.13	N/A	16.89	4.54
MMFed [32]	47.24	21.34	13.97	51.85	61.81	82.34	152.37	N/A	18.09	9.04
FedMultimodal [36]	24.81	17.70	16.40	49.39	62.40	135.08	143.70	N/A	271.32	9.04
FLASH [49]	34.77	22.88	17.44	49.10	62.13	N/A	N/A	N/A	N/A	9.04
Harmony [56]	50.72	21.61	11.41	46.33	63.17	110.24	145.22	N/A	N/A	18.04
Ours w/o Modality Sel.	88.85	71.12	49.80	48.35	66.47	3.90	8.93	N/A	N/A	**0.84**
Ours w/o Client Sel.	90.39	60.54	51.44	53.67	67.41	2.94	13.49	**0.24**	**0.37**	**0.84**
Ours w/o Joint Sel.	89.30	60.22	54.18	**57.43**	66.69	3.86	16.64	**0.24**	1.65	**0.84**
MFedMC (Ours)	**98.87**	**71.28**	**55.09**	53.31	**67.61**	**1.01**	**7.36**	0.48	**0.37**	**0.84**

[1] **MFedMC (Ours)** utilize a consistent configuration with $\delta = 0.2$, $\gamma = 1$, $\alpha_s = 1/3$, $\alpha_c = 1/3$, $\alpha_r = 1/3$ across all datasets.
[2] **Bold** refers to the highest accuracy or the lowest communication overhead.

to isolate the impact of the core algorithmic differences. For method-specific configurations unique to individual baselines, we adopt the recommended settings from their original publications when explicitly specified, or otherwise select the most equitable configuration to ensure fair comparison.

Our MFedMC: For the proposed MFedMC framework, in addition to the configurations mentioned above, these modality encoders do not output logarithmic probabilities but rather provide definitive predicted categories ($\widehat{\mathbb{Y}}$) for the fusion module (ω). We employ Random Forest (RF) as our fusion module, optimizing it to facilitate a more computationally efficient calculation of Shapley values, with the number of trees set at $N_{\omega_k} = 10$. To further reduce computational complexity, we subsampling the dataset for each client k, constructing a sparse background dataset with samples $|D'_k| = 50$ for all clients $k = 1, 2, \ldots, K$, specifically to compute the Shapley values. The sparse background dataset is created through random sampling from the original dataset, ensuring a representative yet significantly smaller subset of the data.

C. Overall Comparison With Baselines

The performance obtained by the proposed MFedMC framework, in comparison with eight baselines on the ActionSense dataset, are presented in Table II and Fig. 4.

Comparison with SOTA Baselines: The proposed MFedMC, along with its three ablation variants employing random selection strategies, demonstrates substantial performance improvements over four SOTA approaches while significantly

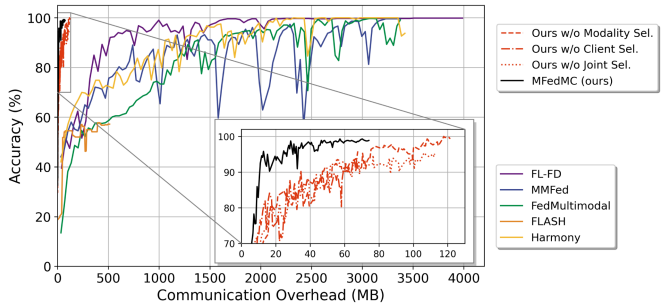


Fig. 4. Accuracy vs. communication overhead on ActionSense dataset under the natural distribution. MFedMC outperforms SOTA baselines with significantly lower communication overhead through decoupled architecture design and strategic joint selection.

reducing communication costs. Notably, even the random selection variants achieve considerable performance gains despite requiring more aggregation iterations due to reduced per-round communication. This accuracy improvement is primarily attributed to our decoupled framework design: clients receive global modality encoders from the server and subsequently perform local fusion module updates. Through server aggregation, modality encoders acquire generalized knowledge from diverse client distributions, while the personalized fusion module fine-tunes locally to optimize multimodal integration based on these updated global encoders. Beyond accuracy improvements, MFedMC reduces communication overhead by nearly an order of magnitude through its selective upload strategy. While random submodel upload strategies [49] can effectively reduce

communication costs to approximately $\frac{1}{M_k+1}$, they lack performance-aware and overhead-aware selection mechanisms, resulting in suboptimal performance-communication trade-offs. In contrast, MFedMC's design is guided by three key takeaways:

i) Different modalities contribute heterogeneously to performance across clients and data distributions.
ii) Selective modality encoder aggregation accelerates convergence compared to aggregating all encoders.
iii) The decoupled architecture enables synergy between generalization (via global encoders) and personalization (via local fusion), which holistic baseline models cannot achieve.

Comparison with Ablation Baselines: Compared to the three ablation baselines implemented through random selection, the proposed MFedMC generally achieves a similar or even lower communication overhead, while outperforming in most scenarios. This is attributed to the selective modality and client communication strategy, where not only does the lower communication cost promote more frequent aggregation, but also the Shapley value and client selection based on local loss can identify the clients with a greater impact in each modality. Due to the inherent randomness, random modality and client selections in some scenarios might marginally outperform MFedMC. This is because randomness, to some extent, ensures that all modalities for all clients have the same probability of being selected. However, this is not the optimal solution as excessive randomness and instability are not preferable.

Comparison between Natural Distribution and IID: The results indicate that IID settings do not always provide clear advantages in model generalization. Both IID and natural distribution settings present unique challenges due to client heterogeneity. In natural distribution scenarios, each client's training and testing data typically share similar yet biased distributions. For example, a hospital in Asia is unlikely to receive patients from North America during deployment. In such cases, the personalized local fusion module adapts effectively to region-specific patterns (e.g., Asian patient characteristics) by fine-tuning based on global modality encoders. Conversely, in IID settings where data is uniformly distributed across clients, the fusion module's personalization capability is underutilized, as local data does not exhibit distinctive patterns requiring specialized adaptation.Furthermore, natural distributions often follow skewed sample distributions. In the PTB-XL dataset, three hospital sites hold 93.54% of the data, with 34 out of 39 sites having fewer than 100 samples each. Similarly, in MELD, six speakers possess 92.68% of the data, with 36 out of 42 speakers having fewer than 100 samples. Clients with limited samples are prone to overfitting, resulting in artificially low local losses that can mislead our client selection strategy. We further investigate the effect of long-tail distribution in Section IV-H.

D. Ablation Study: Modality Selection

1) Trade-Off Analysis: Tables III and IV present ablation results on ActionSense and UCI-HAR without client selection, examining the impact of modality selection parameters $\gamma, \alpha_s, \alpha_c$, and α_r on performance and communication overhead. Optimal performance requires balancing modality informativeness (α_s) and communication cost (α_c). Increasing γ does not always improve accuracy and can substantially increase communication overhead. Selectively uploading only the most informative modalities often yields better performance than uploading all encoders. However, relying solely on Shapley value can lead to a *single modality optimization trap:* if a compact modality contains rich information, it achieves high Shapley values and is repeatedly selected. Server aggregation further refines this encoder, creating a feedback loop where the framework focuses exclusively on one modality while neglecting others. The recency term (α_r) mitigates this by introducing temporal diversity, encouraging the framework to explore different modalities over time rather than prematurely converging to a single dominant modality.

The effectiveness of these selection criteria depends on encoder architecture. In UCI-HAR, both modalities have identical encoder sizes (0.26 MB), making communication

TABLE III
EFFECT OF MODALITY SELECTION WEIGHTS ON ACTIONSENSE: (I) ACCURACY UNDER 25 MB COMMUNICATION CONSTRAINT AND (II) COMMUNICATION OVERHEAD TO ACHIEVE TARGET OF 85% ACCURACY

Method	γ	α_s	α_c	α_r	(i) Accuracy (%) ↑	(ii) Communication Overhead (MB) ↓
FL-FD [31]	-	-	-	-	62.43	48.71
MMFed [32]	-	-	-	-	59.28	82.34
FedMultimodal [36]	-	-	-	-	54.69	135.08
FLASH [49]	-	-	-	-	56.93	N/A
Harmony [56]	-	-	-	-	69.97	110.24
MFedMC ($\delta = 1$)	1	1.0	0.0	0.0	85.14	19.54
		0.0	1.0	0.0	90.91	16.86
		0.0	0.0	1.0	86.34	17.85
		0.0	0.5	0.5	88.95	12.23
		0.5	0.0	0.5	97.92	8.30
		0.5	0.5	0.0	98.05	11.62
		1/3	**1/3**	**1/3**	**99.58**	**3.01**
	2	1.0	0.0	0.0	88.06	20.17
		0.0	1.0	0.0	97.36	21.33
		0.0	0.0	1.0	87.76	14.39
		0.0	0.5	0.5	86.45	22.54
		0.5	0.0	0.5	91.94	14.00
		0.5	0.5	0.0	97.92	13.38
		0.3	0.3	0.3	97.08	4.39
	3	1.0	0.0	0.0	93.45	17.65
		0.0	1.0	0.0	88.97	25.62
		0.0	0.0	1.0	87.91	17.85
		0.0	0.5	0.5	87.28	17.55
		0.5	0.0	0.5	91.66	12.27
		0.5	0.5	0.0	94.44	5.83
		0.3	0.3	0.3	94.71	14.18

TABLE IV
EFFECT OF MODALITY SELECTION WEIGHTS ON UCI-HAR: (I) ACCURACY UNDER 25 MB COMMUNICATION CONSTRAINT AND (II) COMMUNICATION OVERHEAD TO ACHIEVE TARGET OF 60% ACCURACY

Method	γ	α_s	α_c	α_r	(i) Accuracy (%) ↑	(ii) Communication Overhead (MB) ↓
FL-FD [31]	-	-	-	-	71.90	50.13
MMFed [32]	-	-	-	-	43.85	152.37
FedMultimodal [36]	-	-	-	-	37.36	143.70
FLASH [49]	-	-	-	-	48.54	N/A
Harmony [56]	-	-	-	-	45.52	145.22
MFedMC ($\delta = 1$)	**1**	1.0	0.0	0.0	74.51	**35.90**
		0.0	1.0	0.0	37.61	N/A
		0.0	0.0	1.0	56.97	87.57
		0.0	0.5	0.5	57.49	93.70
		0.5	0.0	0.5	74.22	41.16
		0.5	**0.5**	**0.0**	**75.58**	39.41
		0.3	0.3	0.3	75.57	37.66
	2	1.0	0.0	0.0	53.81	103.34
		0.0	1.0	0.0	54.53	99.83
		0.0	0.0	1.0	54.36	92.83
		0.0	0.5	0.5	54.34	96.33
		0.5	0.0	0.5	56.93	N/A
		0.5	0.5	0.0	50.55	96.33
		0.3	0.3	0.3	43.29	N/A

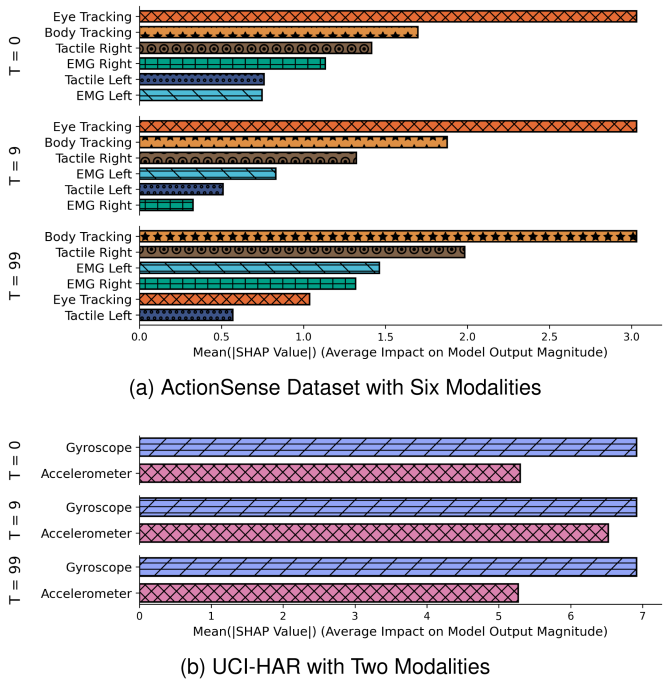


Fig. 5. The mean Shapley value (i.e., impact) of modality encoders throughout the MFedMC iteration. MFedMC demonstrates adaptive modality prioritization: (a) on ActionSense, contributions shift from Eye Tracking (initially dominant due to efficiency) to Body Tracking (information-rich as encoders mature), and (b) on UCI-HAR, Gyroscope exhibits higher impact for capturing rotational movements in activity recognition.

overhead ineffective for differentiation. Moreover, with only two modalities, recency causes cyclic selection, diminishing Shapley-based prioritization. Conversely, MFedMC performs best on datasets like ActionSense with heterogeneous modality dimensions, where diverse encoder sizes enable nuanced selection that effectively balances informativeness, communication cost, and temporal diversity.

2) Analytics on Modality Impact: Beyond guiding modality selection, Shapley values provide an interpretable metric to quantify each modality's contribution to the fusion module's predictions. Fig. 5 illustrates the dynamics of modality impact across communication rounds. In early MFL stages, modalities with simpler features exhibit higher impact as their encoders quickly capture discriminative patterns. As training progresses, the selection dynamics evolve based on the interplay between information richness, encoder complexity, and communication overhead. Modalities that balance informativeness with lower communication costs, such as Body Tracking in ActionSense, emerge as primary contributors, while complex modalities with marginal information gains become less favored despite their potential accuracy improvements.

Fig. 5(a) demonstrates dynamic modality selection on ActionSense. Initially, Eye Tracking dominates due to its simple structure and low communication overhead, enabling frequent uploads and rapid convergence. However, as encoders mature, its contribution diminishes as simpler modalities reach saturation. Body Tracking gradually becomes the primary contributor, offering rich information with moderate communication cost. Tactile Right maintains a secondary position despite high complexity due to its information richness, while Tactile Left remains disadvantaged, likely reflecting limited diversity in left-hand actions. This evolution reflects MFedMC's adaptive prioritization: favoring cost-effective modalities that provide substantial information gains. In contrast, Fig. 5(b) shows subtle changes on UCI-HAR, where both modalities have identical feature dimensions and encoder sizes. Modality selection depends primarily on Shapley values, with recency introducing cyclic patterns that diminish differentiation. The Gyroscope exhibits slightly higher impact than the Accelerometer, as it more effectively captures rotational movements and angular velocities critical for distinguishing activities like running versus climbing stairs.

E. Ablation Study: Client Selection

1) Comparison of Loss-Based Selection Criteria: Recall that our primary objective is to maximize model performance with limited communication resources. In MFL settings, particularly in the MFedMC framework after modality selection, client selection presents a unique challenge: how to compare the quality of different modality encoders across clients without additional assumptions (e.g., access to a public validation dataset or uploading all candidate encoders to the server). Local loss provides a practical solution by enabling cross-modality comparison: a lower loss for a modality encoder indicates more effective learning from local data and suggests higher reliability for global aggregation.

Tables V and VI compare different client selection strategies on the ActionSense and UCI-HAR datasets across various configurations. The results consistently demonstrate that selecting clients with lower local losses outperforms both random selection and selecting clients with higher losses. This finding contrasts with some single-modality FL approaches, such as [48], which propose selecting higher-loss clients to prioritize difficult samples. However, such strategies' objectives fundamentally differ from ours: they prioritize exploration of difficult samples to improve model generalization, whereas our communication-constrained setting prioritizes fast convergence to minimize communication overhead. In our MFedMC framework, convergence speed directly determines the total communication cost required to reach target accuracy. Prioritizing well-trained encoders (i.e., lower-loss clients) ensures stable global aggregation by consistently aggregating high-quality model updates, prevents poorly-trained encoders from degrading the global model, and accelerates convergence to reduce the communication overhead needed.

2) Comparison of Client Selection Frequency: Fig. 6 illustrates the frequency distribution of client selection under the higher-loss and lower-loss strategies, revealing their distinct preferences for certain clients. The higher-loss strategy results in a more uniform distribution, where each client is selected with relatively equal frequency. In contrast, the lower-loss strategy exhibits a clear preference for a subset of clients, leading to a more skewed distribution. This difference stems from the nature of local loss, which reflects not only the information richness of client data but also factors such as client heterogeneity, model optimization dynamics, and data quality. Clients that deviate significantly from the majority due to heterogeneity, high data noise, erroneous labels, or even malicious attacks (e.g., data

TABLE V
EFFECT OF CLIENT SELECTION ON ACTIONSENSE UNDER 5 MB COMMUNICATION CONSTRAINT

δ	γ	α_s	α_c	α_r	Higher Loss: Accuracy (%) ↑	Higher Loss: Communication Overhead (MB) ↓	Higher Loss: Comm. Round	Lower Loss: Accuracy (%) ↑	Lower Loss: Communication Overhead (MB) ↓	Lower Loss: Comm. Round
0.2	1	1.0	0.0	0.0	81.97(+29.16%)	0.05(-80.77%)	99(+80)	98.30(+54.88%)	0.09(-65.48%)	57(+38)
		0.0	1.0	0.0	74.04(-10.00%)	0.06(0.00%)	84(0)	98.87(+20.20%)	0.06(0.00%)	84(0)
		0.0	0.0	1.0	88.17(+3.95%)	0.13(0.00%)	38(0)	84.68(-0.17%)	0.13(0.00%)	38(0)
		0.0	0.5	0.5	85.30(-2.90%)	0.07(0.00%)	76(0)	88.59(+0.84%)	0.07(0.00%)	76(0)
		0.5	0.0	0.5	87.59(-6.27%)	0.05(+1.37%)	92(0)	98.87(+5.80%)	0.06(+12.25%)	83(-9)
		0.5	0.5	0.0	90.62(+5.70%)	0.08(+9.40%)	61(-5)	92.88(+34.04%)	0.14(-64.29%)	35(+23)
		1/3	1/3	1/3	85.06(-8.82%)	0.07(-9.48%)	76(+7)	98.87(+14.52%)	0.06(-15.66%)	79(+13)
	2	1.0	0.0	0.0	82.24(+18.69%)	0.09(-75.98%)	53(+41)	98.89(+6.00%)	0.06(-9.98%)	77(+8)
		0.0	1.0	0.0	82.50(-7.70%)	0.12(0.00%)	41(0)	94.55(+5.77%)	0.12(0.00%)	41(0)
		0.0	0.0	1.0	86.39(-0.95%)	0.26(0.00%)	19(0)	86.11(-1.27%)	0.26(0.00%)	19(0)
		0.0	0.5	0.5	82.55(-5.27%)	0.13(0.00%)	37(0)	89.66(+2.89%)	0.13(0.00%)	37(0)
		0.5	0.0	0.5	87.87(-2.03%)	0.13(-19.20%)	38(+7)	91.48(+2.01%)	0.16(+1.38%)	30(-1)
		0.5	0.5	0.0	86.17(+0.16%)	0.14(+0.68%)	34(-1)	98.60(+14.61%)	0.14(+0.59%)	35(0)
		1/3	1/3	1/3	86.47(-2.82%)	0.13(-1.71%)	38(+1)	97.05(+9.07%)	0.14(+5.67%)	35(-2)
	3	1.0	0.0	0.0	86.75(+18.81%)	0.17(-68.80%)	30(+21)	89.25(+22.24%)	0.21(-61.33%)	24(+15)
		0.0	1.0	0.0	79.59(-7.19%)	0.18(0.00%)	27(0)	93.30(+8.80%)	0.18(0.00%)	27(0)
		0.0	0.0	1.0	86.77(+5.03%)	0.39(0.00%)	12(0)	86.67(+4.90%)	0.39(0.00%)	12(0)
		0.0	0.5	0.5	83.65(-7.01%)	0.19(0.00%)	25(0)	91.21(+1.41%)	0.19(0.00%)	25(0)
		0.5	0.0	0.5	89.55(-2.39%)	0.21(-2.11%)	24(+1)	90.20(-1.68%)	0.26(+20.91%)	19(-4)
		0.5	0.5	0.0	81.83(-4.73%)	0.18(-9.07%)	27(+3)	93.00(+8.28%)	0.21(+2.48%)	24(0)
		1/3	1/3	1/3	81.97(-7.74%)	0.19(-2.31%)	26(0)	93.44(+5.16%)	0.21(+8.09%)	24(-2)

Improvement over Baseline (Random Client Selection)
$\geq$ 10%
0% to 10%
-10% to 0%
$\leq$ -10%

TABLE VI
EFFECT OF CLIENT SELECTION ON UCI-HAR UNDER 5 MB COMMUNICATION CONSTRAINT

δ	γ	α_s	α_c	α_r	Higher Loss: Accuracy (%) ↑	Higher Loss: Communication Overhead (MB) ↓	Higher Loss: Comm. Round	Lower Loss: Accuracy (%) ↑	Lower Loss: Communication Overhead (MB) ↓	Lower Loss: Comm. Round
0.2	1	1.0	0.0	0.0	47.78(-22.79%)	0.05(0.00%)	95(0)	68.77(+11.13%)	0.05(0.00%)	95(0)
		0.0	1.0	0.0	33.80(-39.88%)	0.05(0.00%)	95(0)	55.77(-0.80%)	0.05(0.00%)	95(0)
		0.0	0.0	1.0	40.31(-18.33%)	0.05(0.00%)	95(0)	68.27(+38.30%)	0.05(0.00%)	95(0)
		0.0	0.5	0.5	37.29(-23.51%)	0.05(0.00%)	95(0)	67.60(+38.67%)	0.05(0.00%)	95(0)
		0.5	0.0	0.5	42.39(-34.28%)	0.05(0.00%)	95(0)	70.62(+9.50%)	0.05(0.00%)	95(0)
		0.5	0.5	0.0	47.70(-14.81%)	0.05(0.00%)	95(0)	65.16(+16.36%)	0.05(0.00%)	95(0)
		1/3	1/3	1/3	39.74(-24.53%)	0.05(0.00%)	95(0)	71.28(+35.37%)	0.05(0.00%)	95(0)
	2	1.0	0.0	0.0	36.74(-22.25%)	0.11(0.00%)	47(0)	65.77(+39.19%)	0.11(0.00%)	47(0)
		0.0	1.0	0.0	35.43(-29.30%)	0.11(0.00%)	47(0)	67.37(+34.43%)	0.11(0.00%)	47(0)
		0.0	0.0	1.0	41.61(-15.11%)	0.11(0.00%)	47(0)	68.22(+39.17%)	0.11(0.00%)	47(0)
		0.0	0.5	0.5	37.24(-27.29%)	0.11(0.00%)	47(0)	65.35(+27.58%)	0.11(0.00%)	47(0)
		0.5	0.0	0.5	39.52(-23.88%)	0.11(0.00%)	47(0)	66.99(+29.02%)	0.11(0.00%)	47(0)
		0.5	0.5	0.0	37.60(-20.59%)	0.11(0.00%)	47(0)	69.74(+47.28%)	0.11(0.00%)	47(0)
		1/3	1/3	1/3	36.02(-26.67%)	0.11(0.00%)	47(0)	67.83(+38.10%)	0.11(0.00%)	47(0)

Improvement over Baseline (Random Client Selection)
$\geq$ 10%
0% to 10%
-10% to 0%
$\leq$ -10%

poisoning) tend to exhibit higher local losses. Consequently, while the lower-loss strategy may favor clients with more homogeneous data, the higher-loss strategy is more susceptible to selecting clients that deviate from the main client data distribution.

F. Effect of Class and Modality non-IID Settings

Class Non-IID Setting: To simulate class non-IID setting, we redistribute client datasets through a Dirichlet distribution Dir(β) in the ActionSense dataset, with results shown in Fig. 7(a). It can be observed that our method consistently outperforms all baselines at any β value, and the impact of non-IIDness on our method is minimal, whether for our full approach or our ablation baselines with random selection. This advantage primarily stems from our strategy of separating modality encoders and fusion modules while keeping fusion modules local as personalization implementations. This optimizes fusion modules to learn local data distributions without interference from other clients' data distributions. The baseline methods lack such unique personalization configurations, resulting in slower convergence under non-IID conditions.

Modality Non-IID Setting: To simulate scenarios where clients do not possess all modalities (i.e., have missing modalities), we evaluate the performance of our system by randomly removing modalities from clients at certain probabilities, as illustrated in Fig. 7(b). Thus, a missing modality rate of 0% corresponds to the case shown in Fig. 4. A missing modality rate of 80% corresponds to the case where each client has at least two modalities (i.e., $M_k \geq 2$ for all clients k). Also, recall that even with the missing modality rate of 0%, the ActionSense dataset inherently has missing modalities, with subjects 06 through 09 missing both left and right tactile data. We see that our MFedMC framework still demonstrates superior performance compared to baselines across missing modalities. Under extreme modality non-IID (missing modality rate of 80%), the MFedMC framework inevitably experiences performance degradation due to the significant loss of information from the missing modalities. Interestingly, the performance of the MFedMC framework under extreme modality non-IID (missing modality rate of 80%) still surpasses that of the baselines under no modality non-IID (missing modality rate of 0%). This underscores the advantages provided by MFedMC: the fusion module enhances accuracy through personalization, while joint selection identifies the most

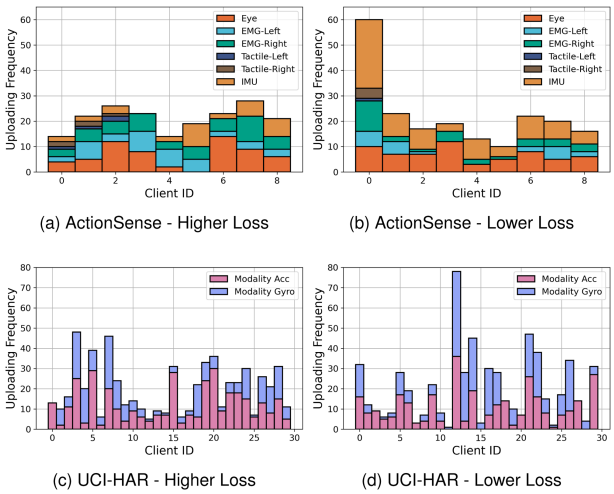


Fig. 6. Histograms of the client selection frequency based on the criteria of higher and lower loss.

informative modalities, optimizes client models, and reduces communication overhead.

G. Effect of Heterogeneous Network

In practice, different clients possess varying available network bandwidth, inhibiting their ability to upload all model parameters. Through modality selection, MFedMC can adaptively select which modality encoders to transmit based on bandwidth limitations, ensuring efficient use of available communication channels and minimizing bottlenecks due to bandwidth constraints. To demonstrate this, we conduct experiments in a setting where clients have different uplink communication restrictions for uploading modality encoders to the server, due to heterogeneous transmit powers and uplink channels. Specifically, using the ActionSense dataset, we classify the clients into three levels based on their communication capabilities:

- *Unrestricted Communication:* Clients 1-2 are able to upload all modality encoders.
- *Moderate Communication Restrictions:* Clients 3-5 are able to upload the Eye Tracking, EMG Left, EMG Right, and Body Tracking models.
- *Severe Communication Restrictions:* Clients 6-9 are limited to uploading the Eye Tracking, EMG Left, and EMG Right models due to their manageable sizes.

For the SOTA baselines, FLASH [49] and Harmony [56] allow for the segmentation of models, enabling each client to participate in the FL process. In contrast, the other three baselines, FL-FD [31], MMFed [32], and FedMultimodal [36], utilize end-to-end models, which means that under communication restrictions, only clients 1-2 are able to engage in FL.

Fig. 8 shows the result of the proposed MFedMC and eight baselines in the heterogeneous network scenario. Compared to the homogeneous communication scenario depicted in Fig. 4, we see that while communication constraints lead to a drop in accuracy in the early stages of the MFedMC framework, our methodology ultimately reaches roughly the same accuracy. This demonstrates the capability of the MFedMC framework to handle heterogeneous network conditions, ensuring that all clients, regardless of their individual restrictions, contribute to and benefit from the FL process. It is noteworthy that most SOTA baseline methods are not designed to operate effectively under heterogeneous network conditions. If clients have varying network bandwidths, which is common, the majority of clients (in our setup, Clients 3-9) will be unable to upload complete models to the server, thereby being unable to contribute to FL due to bandwidth limitations. As shown in Fig. 8, this limitation causes some SOTA baselines (FL-FD, MMFed, and FedMultimodal) to converge to a low accuracy level, reflecting the maximum performance achievable using data from only Clients 1-2. Other SOTA baselines (FLASH and Harmony) allow uploading parts of the complete model; thus, although they can continue to improve accuracy, the improvement is much slower compared to our MFedMC algorithm that selects modalities and clients. In contrast, our modality selection strategically prioritizes and optimizes data transmission, enabling all clients to continuously participate in the FL network without being hindered by their individual bandwidth restrictions.

H. Effect of Long-Tail Distribution

To investigate the robustness of MFedMC under imbalanced client data distributions, particularly considering that lower loss-based client selection might favor clients with smaller or simpler datasets, we conduct experiments with varying degrees of long-tail distribution across clients. Following prior work [57], we control the client sample distribution through the Imbalance Factor (IF), where higher IF values represent more severe long-tail distributions with greater disparity in sample sizes across clients. Furthermore, we explore whether incorporating the recency term from modality selection into client selection could mitigate potential bias caused by loss-based selection. We compare five client selection strategies with different loss-recency weight combinations: pure loss-based selection, hybrid strategies that balance loss and recency considerations, and pure recency-based selection that prioritizes clients not recently selected.

As shown in Fig. 9, several key observations emerge from our experiments. First, pure lower loss-based selection (weight configuration $(1.0, 0.0)$) maintains acceptable performance across all imbalance levels, remaining competitive even under extreme conditions (IF=100). This suggests that while selecting clients solely based on lower local loss may have theoretical limitations, its practical impact appears more limited than initially anticipated. Moreover, the rationale for lower loss-based selection is fundamentally aligned with MFedMC's core objective of communication efficiency. In communication-constrained FL, convergence speed directly determines the total communication cost required to reach target accuracy. By prioritizing clients with lower losses, this selection strategy consistently aggregates encoder updates, thereby accelerating convergence and reducing the communication overhead needed to achieve satisfactory performance. Second, while incorporating recency provides some improvements in certain scenarios (e.g., $(0.2, 0.8)$ performs best at IF=20, 50, 100), the benefits are limited and inconsistent. Compared with pure lower loss-based selection, improvements are often marginal, optimal configurations vary significantly

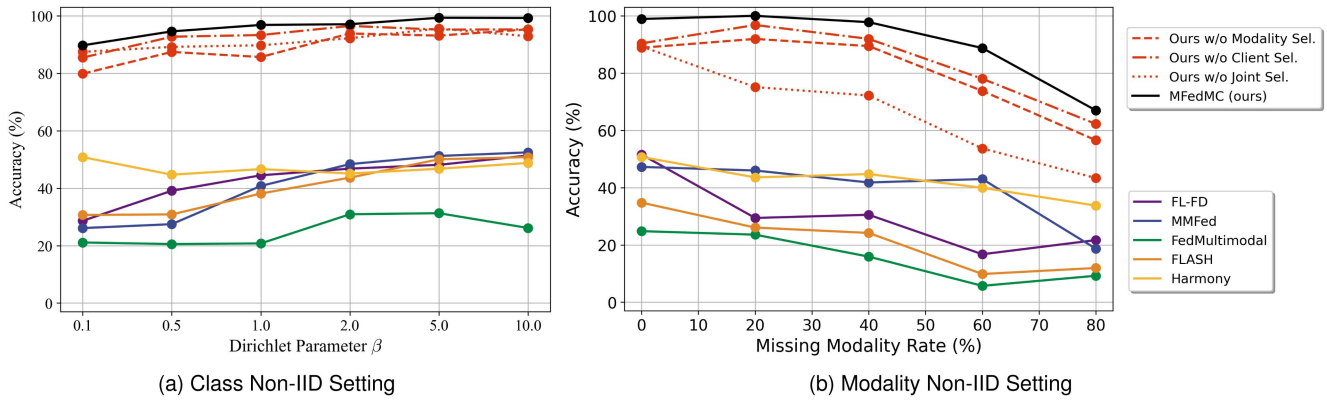


Fig. 7. Effect of class and modality non-IID settings on ActionSense dataset. MFedMC demonstrates robust performance under both non-IID conditions: (a) maintaining superior accuracy across varying Dirichlet parameters through local fusion module personalization, and (b) outperforming baselines under extreme missing modality rates.

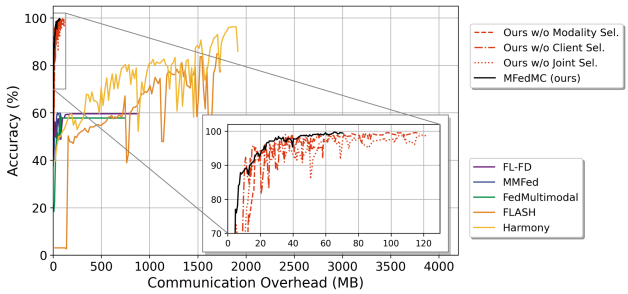


Fig. 8. Effect of network heterogeneity. MFedMC enables all clients to participate regardless of bandwidth constraints through strategic modality and client selection, achieving comparable final accuracy to homogeneous settings while most baselines either converge to low accuracy or exhibit significantly slower improvement due to inability to handle heterogeneous network conditions.

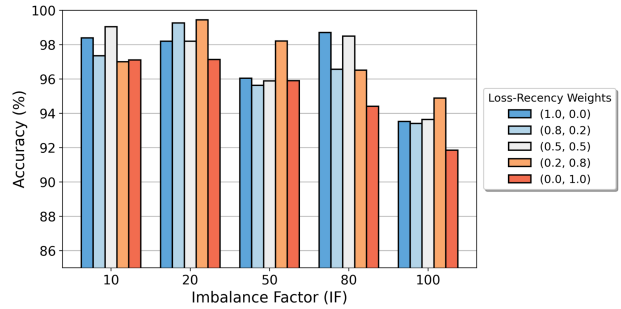


Fig. 9. Effect of imbalance data distribution and loss-recency weight combinations. Pure loss-based selection (1.0, 0.0) maintains competitive performance across all imbalance levels, while hybrid strategies incorporating recency show marginal improvements with varying optimal configurations across different imbalance factors.

across imbalance levels (e.g., $(0.2, 0.8)$ degrades at IF=10, 80), and the performance gap remains small across all tested factors.

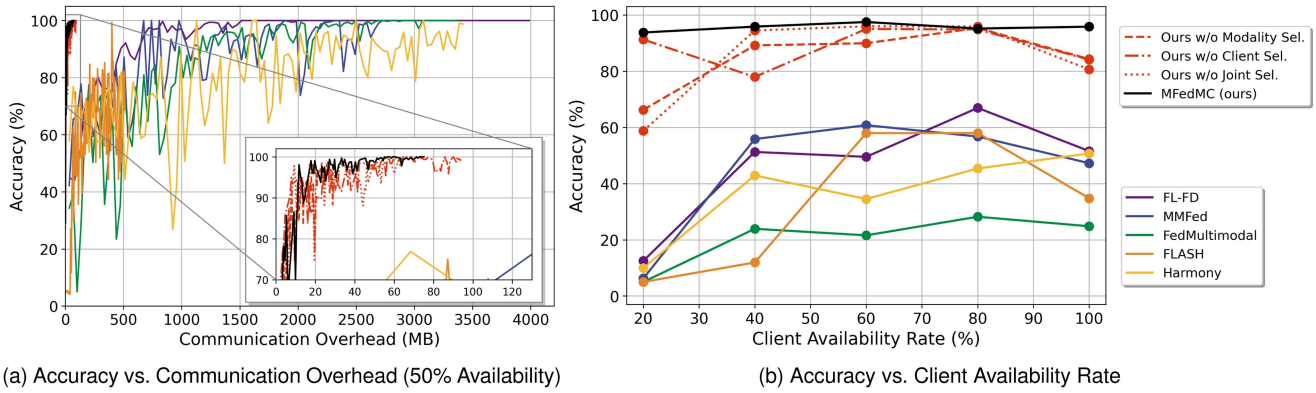
These findings indicate that while incorporating recency has conceptual merit, it introduces hyperparameter sensitivity without consistent improvements and may compromise communication efficiency. Notably, the role of recency in client selection fundamentally differs from its role in modality selection. In modality selection, incorporating recency to diversify modality choices is beneficial because multiple modality encoders jointly contribute to predictions through the fusion module, which can adaptively weigh and compensate for weaker modality predictions. In contrast, client selection aims to identify clients with the best-trained modality encoders for global aggregation, where selecting poorly-trained encoders (even for diversity) directly slows convergence and increases the communication overhead required to reach target accuracy. Therefore, we maintain the pure lower loss-based client selection strategy, which prioritizes high-quality encoders to maximize aggregation effectiveness and communication efficiency.

I. *Effect of Client Dynamics and Availability*

Fig. 10 demonstrates the robustness of our framework under client dynamics caused by stragglers and client churn, where MFedMC and baselines can only perform federated training on a subset of available clients. We observe that MFedMC and its ablation variants maintain superior performance even when a large proportion of clients are unavailable. In contrast, client availability has a more pronounced impact on SOTA baselines, particularly under lower availability rates. This is because when client availability decreases, baseline methods struggle to acquire sufficient global representations as they require all clients to upload complete models in each round. Our method, however, not only obtains richer global representations through reduced communication overhead (enabling more frequent aggregation rounds), but also compensates for incomplete global information through the personalized fusion module strategy. Specifically, the local fusion module can adaptively weight and combine the available modality encoders based on local data characteristics, effectively mitigating the negative impact of missing updates from unavailable clients. This design makes MFedMC naturally resilient to client churn and stragglers, as the personalized fusion module serves as a robust buffer that maintains performance stability even when global modality encoders receive updates from only a subset of clients.

J. *Integration of Communication Compression*

Our MFedMC framework reduces communication overhead through joint modality and client selection, and also can integrate



(a) Accuracy vs. Communication Overhead (50% Availability) (b) Accuracy vs. Client Availability Rate

Fig. 10. Effect of client availability. MFedMC achieves faster convergence with lower communication overhead and maintains superior accuracy across different client availability rates compared to baseline methods.

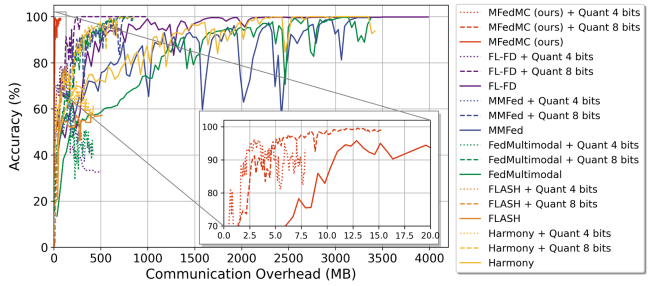


Fig. 11. Effect of quantization precision. MFedMC is compatible with different quantization levels (4-bit and 8-bit) and consistently achieves faster convergence with lower communication overhead compared to baseline methods.

TABLE VII
END-TO-END SYSTEM-LEVEL RUNTIME COMPARISON

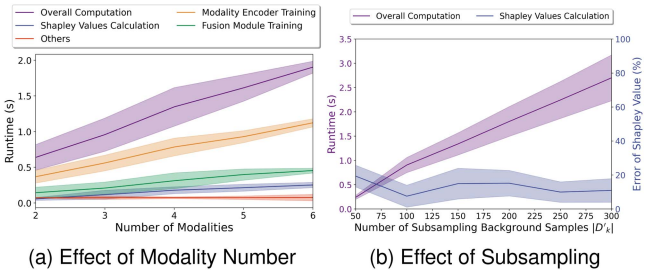
Method	Training Time (mins)	Communication Time (mins)	Total Time (mins)
FL-FD [31]	31.11	100.29	131.40
MMFed [32]	27.03	130.15	157.18
FedMultimodal [36]	29.44	130.36	159.80
FLASH [49]	27.77	130.15	157.91
Harmony [56]	37.60	130.15	167.75
Ours w/o Modality Sel.	19.52	1.75	21.27
Ours w/o Client Sel.	23.57	1.63	25.20
Ours w/o Joint Sel.	19.13	1.87	21.00
MFedMC (Ours)	23.56	2.10	25.67

with communication compression algorithms to further minimize communication costs and bandwidth requirements. As illustrated in Fig. 11, comparative analysis reveals that our method consistently outperforms baseline approaches when combined with quantization techniques, demonstrating accelerated convergence and reduced communication costs across all precision levels. Particularly noteworthy is the performance differentiation at varying quantization levels: while all methods maintain normal convergence at 8-bit quantization, only our approach sustains acceptable accuracy at 4-bit quantization, whereas baseline methods fail to converge entirely. This robust performance can be attributed to our personalized local fusion module, which effectively mitigates quantization errors. Conversely, baseline methods amplify these errors as they propagate through the fusion module directly connected to the task head. Consequently, the distinctive architecture of localized fusion modules within the MFedMC framework demonstrates remarkable compatibility with quantization-induced perturbations, enabling effective integration with additional communication compression algorithms to achieve further reductions in transmission overhead.

K. Computational Efficiency Analysis

End-to-End System-Level Time Analysis: We measure the actual wall-clock training time for all methods on the ActionSense dataset over 100 communication rounds. The total time comprises two components: (1) *Training time:* the actual wall-clock time for local model training, and (2) *Communication time:* the simulated transmission time based on realistic network conditions with an uplink bandwidth of 10 Mbps (representative of typical edge/IoT devices), a protocol encoding overhead of 1.2×, and a forward error correction (FEC) overhead of 1.5×, calculated as $T_{\text{comm}} = \text{Model Size (bytes)} \times 1.2 \times 1.5 \div (10 \times 10^6/8)$ seconds. As shown in Table VII, MFedMC achieves a comparable or lower training time relative to baselines (23.56 mins vs. 27-37 mins) while demonstrating a significant reduction in communication time (2.10 mins vs. 100-130 mins). Although MFedMC requires slightly more training time than some of our ablation variants due to Shapley value computation, this modest computational overhead is well justified by the accuracy improvements (4-10% as shown in Table II) and the 50-60× reduction in communication time. These results confirm that communication is the dominant bottleneck, accounting for 76-84% of the total time for baseline methods. Our joint selection strategy effectively addresses this challenge, achieving an overall 5-6× speedup in end-to-end training time.

Shapley Value Computation Overhead: Fig. 12(a) illustrates the impact of varying the number of modalities on runtime performance. The computational overhead of Shapley value calculation is independent of the number of parameters in the modality encoders. According to the computational complexity of $\mathcal{O}(N_{\omega_k} L_{\omega_k} H_{\omega_k} |D'_k|)$ derived in Sec. III-D, the runtime associated with Shapley value computation remains independent of the number of modalities, exhibiting only a marginal increase attributable to auxiliary data processing operations such as normalization. Relative to the dominant computational components, modality encoder training and fusion module



(a) Effect of Modality Number (b) Effect of Subsampling

Fig. 12. Runtime analysis of Shapley value computation. Solid lines represent mean values, while shaded regions indicate standard deviation.

training, the runtime for Shapley value calculation is substantially lower, demonstrating that our approach introduces no significant scalability challenges across varying numbers of modalities. Fig. 12(b) presents the relationship between subsampling size and both runtime performance and Shapley value estimation error. As the number of subsampled points $|D'_k|$ increases, runtime exhibits linear growth consistent with the complexity bound $\mathcal{O}(N_{\omega_k} L_{\omega_k} H_{\omega_k} |D'_k|)$. Notably, the estimation error decreases only marginally from 19.33% to 10.91% as the background sample size increases from $|D'_k| = 50$ to $|D'_k| = 300$, demonstrating a favorable trade-off between computational efficiency and approximation fidelity.

V. Conclusion

In this paper, we introduced the MFedMC framework that minimizes communication overhead and enhances learning efficiency through joint modality and client selection strategies. By optimizing modality selection with Shapley values, modality encoder sizes, and recency, as well as favoring clients with lower local loss for client selection, we achieved considerable recognition accuracy and up to 20× increase in communication efficiency. The proposed MFedMC is flexible, suitable for heterogeneous clients, offers modular modality encoders that can be detached, and provides impact assessment for data modalities. Extensive experiments across a diverse range of real-world datasets, including wearable sensors, healthcare, NLP, and remote sensing satellite datasets, demonstrated the outstanding performance of the MFedMC framework across various data modalities.

Our future work will focus on enhancing the adaptability of MFedMC through dynamic configuration. For modality selection, we could adjust the weight of communication overhead based on the dynamically available communication resources in the real world (such as higher bandwidth at night), allowing for the uploading of more modality encoders. For client selection, we may consider a dynamic strategy based on local loss, employing selection of higher local losses at the initial stages of FL to speed up convergence, and transitioning to the proposed lower local loss selection towards the end to optimize local minima.

References

[1] L. Yuan, D.-J. Han, V. P. Chellapandi, S. H. Zak, and C. G. Brinton, "FedMFs: Federated multimodal fusion learning with selective modality communication," in *Proc. IEEE Int. Conf. Commun.*, 2024, pp. 287–292.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[3] L. Yuan, Z. Wang, L. Sun, P. S. Yu, and C. G. Brinton, "Decentralized federated learning: A survey and perspective," *IEEE Internet Things J.*, vol. 11, no. 21, pp. 34617–34638, Nov. 2024.

[4] L. Yuan, Z. Wang, and C. G. Brinton, "Digital ethics in federated learning," *IEEE Internet Comput.*, vol. 28, no. 5, pp. 66–74, Sep./Oct. 2024.

[5] W. Fang, D.-J. Han, L. Yuan, S. Hosseinalipour, and C. G. Brinton, "Federated sketching LoRA: A flexible framework for heterogeneous collaborative fine-tuning of LLMs," 2025, *arXiv:2501.19389*.

[6] H. Yu et al., "FedHAR: Semi-supervised online learning for personalized federated human activity recognition," *IEEE Trans. Mobile Comput.*, vol. 22, no. 6, pp. 3318–3332, Jun. 2023.

[7] Q. Wu, X. Chen, Z. Zhou, and J. Zhang, "FedHome: Cloud-edge based personalized federated learning for in-home health monitoring," *IEEE Trans. Mobile Comput.*, vol. 21, no. 8, pp. 2818–2832, Aug. 2022.

[8] S. Baghersalimi, T. Teijeiro, A. Aminifar, and D. Atienza, "Decentralized federated learning for epileptic seizures detection in low-power wearable systems," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 6392–6407, May 2024.

[9] Y.-M. Lin, Y. Gao, M.-G. Gong, S.-J. Zhang, Y.-Q. Zhang, and Z.-Y. Li, "Federated learning on multimodal data: A comprehensive survey," *Mach. Intell. Res.*, vol. 20, no. 4, pp. 539–553, 2023.

[10] J. Chen and A. Zhang, "FedMBridge: Bridgeable multimodal federated learning," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 7667–7686.

[11] L. Yuan, Y. Ma, L. Su, and Z. Wang, "Peer-to-peer federated continual learning for naturalistic driving action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5249–5258.

[12] V. P. Chellapandi, L. Yuan, S. H. Zak, and Z. Wang, "A survey of federated learning for connected and automated vehicles," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst.*, 2023, pp. 2485–2492.

[13] V. P. Chellapandi, L. Yuan, C. G. Brinton, S. H. Zak, and Z. Wang, "Federated learning for connected and automated vehicles: A survey of existing approaches and challenges," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 119–137, Jan. 2024.

[14] G. Lee, S.-H. Jung, and D. Han, "An adaptive sensor fusion framework for pedestrian indoor navigation in dynamic environments," *IEEE Trans. Mobile Comput.*, vol. 20, no. 2, pp. 320–336, Feb. 2021.

[15] Y. Li, L. Liu, H. Qin, S. Deng, M. A. El-Yacoubi, and G. Zhou, "Adaptive deep feature fusion for continuous authentication with data augmentation," *IEEE Trans. Mobile Comput.*, vol. 22, no. 10, pp. 5690–5705, Oct. 2023.

[16] J. Zhang, Q. Wang, Q. Wang, and Z. Zheng, "Multimodal fusion framework based on statistical attention and contrastive attention for sign language recognition," *IEEE Trans. Mobile Comput.*, vol. 23, no. 2, pp. 1431–1443, Feb. 2024.

[17] P. Kairouz et al., "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, no. 1–2, pp. 1–210, 2021.

[18] Y. Zhao, P. Barnaghi, and H. Haddadi, "Multimodal federated learning on IoT data," in *Proc. IEEE/ACM Seventh Int. Conf. Internet-of-Things Des. Implementation*, 2022, pp. 43–54.

[19] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, Sep. 2020.

[20] W. Y. B. Lim et al., "Federated learning in mobile edge networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 2031–2063, thirdquarter 2020.

[21] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amini, "A survey on federated learning for resource-constrained IoT devices," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 1–24, Jan. 2022.

[22] W. Fang, D.-J. Han, and C. G. Brinton, "Submodel partitioning in hierarchical federated learning: Algorithm design and convergence analysis," in *Proc. IEEE Int. Conf. Commun.*, 2024, pp. 268–273.

[23] J. Zhang et al., "FedAda: Fast-convergent adaptive federated learning in heterogeneous mobile edge computing environment," in*World Wide Web*, vol. 25, no. 5, pp. 1971–1998, 2022.

[24] Y.-W. Chu et al., "Mitigating biases in student performance prediction via attention-based personalized federated learning," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, 2022, pp. 3033–3042.

[25] A. M. Abdelmoniem, C.-Y. Ho, P. Papageorgiou, and M. Canini, "Empirical analysis of federated learning in heterogeneous environments," in *Proc. 2nd Eur. Workshop Mach. Learn. Syst.*, 2022, pp. 1–9.

[26] L. S. Shapley et al., "A value for N-Person games," in *Contributions to the Theory of Games*, vol. II, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton Univ. Press, 1953, pp. 307–317.

[27] H. P. Young, "Monotonic solutions of cooperative games," *Int. J. Game Theory*, vol. 14, no. 2, pp. 65–72, 1985.
[28] M. Sundararajan and A. Najmi, "The many shapley values for model explanation," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9269–9278.
[29] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.
[30] S. M. Lundberg et al., "From local explanations to global understanding with explainable ai for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 2522–5839, 2020.
[31] P. Qi, D. Chiaro, and F. Piccialli, "FL-FD: Federated learning-based fall detection with multimodal data fusion," *Inf. Fusion*, vol. 99, 2023, Art. no. 101890.
[32] B. Xiong, X. Yang, F. Qi, and C. Xu, "A unified framework for multi-modal federated learning," *Neurocomputing*, vol. 480, pp. 110–118, 2022.
[33] S. Chen and B. Li, "Towards optimal multi-modal federated learning on non-IID data with hierarchical gradient blending," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2022, pp. 1469–1478.
[34] J. Chen and A. Zhang, "FedMSplit: Correlation-adaptive federated multi-task learning across multimodal split networks," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2022, pp. 87–96.
[35] G. Bao, Q. Zhang, D. Miao, Z. Gong, and L. Hu, "Multimodal federated learning with missing modality via prototype mask and contrast," 2023, *arXiv:2312.13508*.
[36] T. Feng et al., "FedMultimodal: A benchmark for multimodal federated learning," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2023, pp. 4035–4045.
[37] H. Q. Le, C. M. Thwal, Y. Qiao, Y. L. Tun, M. N. Nguyen, and C. S. Hong, "Cross-modal prototype based multimodal federated learning under severely missing modality," *Inf. Fusion*, vol. 122, p. 103219, 2025.
[38] T. Zheng, A. Li, Z. Chen, H. Wang, and J. Luo, "AutoFed: Heterogeneity-aware federated multimodal learning for robust autonomous driving," in *Proc. 29th Annu. Int. Conf. Mobile Comput. Netw.*, 2023, pp. 1–15.
[39] Q. Yu, Y. Liu, Y. Wang, K. Xu, and J. Liu, "Multimodal federated learning via contrastive representation ensemble," in *Proc. 11th Int. Conf. Learn. Representations*, Kigali, Rwanda, 2023.
[40] B. Xiong, X. Yang, Y. Song, Y. Wang, and C. Xu, "Client-adaptive cross-model reconstruction network for modality-incomplete multimodal federated learning," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 1241–1249.
[41] L. Yuan, D.-J. Han, S. Wang, and C. Brinton, "Local-cloud inference offloading for LLMs in multi-modal, multi-task, multi-dialogue settings," in *Proc. 26th Int. Symp. Theory, Algorithmic Found., Protoc. Des. Mobile Netw. Mobile Comput.*, 2025, pp. 201–210.
[42] Y. Xu, Z. Jiang, H. Xu, Z. Wang, C. Qian, and C. Qiao, "Federated learning with client selection and gradient compression in heterogeneous edge systems," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 5446–5461, May 2024.
[43] B. Luo, W. Xiao, S. Wang, J. Huang, and L. Tassiulas, "Tackling system and statistical heterogeneity for federated learning with adaptive client sampling," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2022, pp. 1739–1748.
[44] L. Yuan, L. Su, and Z. Wang, "Federated transfer-ordered-personalized learning for driver monitoring application," *IEEE Internet Things J.*, vol. 10, no. 20, pp. 18292–18301, Oct. 2023.
[45] L. Fu, H. Zhang, G. Gao, M. Zhang, and X. Liu, "Client selection in federated learning: Principles, challenges, and opportunities," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 21811–21819, Dec. 2023.
[46] S. Wang, R. Morabito, S. Hosseinalipour, M. Chiang, and C. G. Brinton, "Device sampling and resource optimization for federated learning in cooperative edge networks," *IEEE/ACM Trans. Netw.*, vol. 32, no. 5, pp. 4365–4381, 2024.
[47] Q. Pan, H. Cao, Y. Zhu, J. Liu, and B. Li, "Contextual client selection for efficient federated learning over edge devices," *IEEE Trans. Mobile Comput.*, vol. 23, no. 6, pp. 6538–6548, Jun. 2024.
[48] Y. J. Cho, J. Wang, and G. Joshi, "Client selection in federated learning: Convergence analysis and power-of-choice selection strategies," 2020, *arXiv:2010.01243*.
[49] B. Salehi, J. Gu, D. Roy, and K. Chowdhury, "FLASH: Federated learning for automated selection of high-band mmWave sectors," in *Proc. IEEE Conf. Comput. Commun.*, 2022, pp. 1719–1728.
[50] J. DelPreto et al., "ActionSense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 13800–13813.
[51] D. Anguita et al., "A public domain dataset for human activity recognition using smartphones," *Esann*, vol. 3, no. 1, pp. 3–4, 2013.
[52] P. Wagner et al., "PTB-XL, a large publicly available electrocardiography dataset," *Sci. Data*, vol. 7, no. 1, 2020, Art. no. 154.
[53] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep learning for ECG analysis: Benchmarks and insights from PTB-XL," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1519–1528, May 2021.
[54] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 527–536.
[55] C. Persello et al., "2023 IEEE GRSS data fusion contest: Large-scale fine-grained building classification for semantic urban reconstruction," 2022, doi: 10.21227/mrnt-8w27.
[56] X. Ouyang et al., "Harmony: Heterogeneous multi-modal federated learning through disentangled model training," in *Proc. 21st Annu. Int. Conf. Mobile Syst., Appl. Serv.*, 2023, pp. 530–543.

Liangqi Yuan (Graduate Student Member, IEEE) received the BE degree in photo-electronic information science and engineering from Beijing Information Science and Technology University, Beijing, China, in 2020, and the MS degree in electrical and computer engineering from Oakland University, Rochester, MI, USA, in 2022. He is currently working toward the PhD degree in electrical and computer engineering with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. His research interests include multimodal learning, mobile computing, and machine learning.

Dong-Jun Han (Member, IEEE) received the BS degree in mathematics and electrical engineering and the MS and PhD degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2016, 2018, and 2022, respectively. He is currently an assistant professor with the Department of Computer Science and Engineering, Yonsei University, South Korea. His research interests include the intersection of communications, networking, and machine learning, and specifically in distributed/federated machine learning and network optimization.

Su Wang received the BS (with Distinction) and PhD degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 2018 and 2023, respectively. He is currently a joint lecturer and postdoctoral research associate with the School of Electrical and Computer Engineering, Princeton University.

Devesh Upadhyay (Senior Member, IEEE) received the MS and PhD degrees in mechanical engineering from The Ohio State University, Columbus, USA. He was a senior technical leader with Ford Research, where he led the Core AI/ML Quantum Computing Team. He is currently the technical director for AI/ML and autonomy with Saab Inc.

Christopher G. Brinton (Senior Member, IEEE) received the MSc and PhD degrees in electrical engineering from Princeton University, in 2013 and 2016, respectively. He is currently the elmore associate professor of ECE with Purdue University. He was the recipient of four of the U.S. top early career awards, from the National Science Foundation (CAREER), Office of Naval Research (YIP), Defense Advanced Research Projects Agency (YFA), and Air Force Office of Scientific Research (YIP), Intel Rising Star Faculty Award, and Qualcomm Faculty Award.