Staleness-Aware Pruning and Resource Allocation for Wireless Multimodal Federated Learning

Chunfeng Xie, Zhixiong Chen, *Member, IEEE*, Wenqiang Yi, *Member, IEEE*, Chan-Tong Lam, *Senior Member, IEEE*, and Arumugam Nallanathan, *Fellow, IEEE*

***Abstract*—Multimodal federated learning (MFL) allows distributed edge devices to learn from heterogeneous sensing modalities without sharing raw data. Its wireless deployment remains challenging because clients may have missing modalities, heterogeneous computing capabilities and channel conditions, and costly multimodal model updates. We propose mFedDNP, a communication- and computation-efficient framework integrating modal compensation, staleness-aware dynamic pruning, client scheduling, and radio-resource allocation. In each round, the server selects clients and constructs pruned unimodal submodels tailored to modality availability, computing capability, and uplink conditions. Historical unimodal gradients compensate for unavailable modalities, while discrepancy-based updates avoid storing a full client-modality gradient array at the server. For a multi-antenna orthogonal frequency-division multiple access (OFDMA) uplink, we jointly optimize client participation, pruning masks, and resource block (RB) allocation under a round-latency constraint. A convergence upper bound explicitly characterizes cumulative gradient staleness and yields a tractable per-round objective. The resulting policy applies staleness-ranked pruning and reduces client-RB assignment to a maximum weight bipartite matching problem. Experiments on CREMA-D and UCI-HAR show that mFedDNP improves Macro-F1 by up to 4.89 percentage points and reduces the time to target performance by up to 23.52% over representative baselines, demonstrating an improved accuracy and latency tradeoff under modality and resource heterogeneity.**

***Index Terms*—Multimodal federated learning, beamforming, dynamic neural network pruning, modal compensation**

I. Introduction

The rapid proliferation of edge devices equipped with heterogeneous sensors, such as cameras, microphones, and inertial units, is enabling a wide range of intelligent applications, including emotion recognition, human activity recognition, healthcare monitoring, and immersive human-machine interaction [2]–[4]. Compared with unimodal learning, multimodal learning can exploit complementary information across different sensing sources and thus often achieves higher robustness and better inference accuracy. However, centralized training of multimodal models requires collecting large amounts of privacy-sensitive data at a remote server, which incurs substantial communication overhead and raises serious privacy concerns. Federated learning (FL) [5] offers an attractive alternative by allowing distributed clients to collaboratively train a shared model without uploading raw data. As a result, multimodal federated learning (MFL) has emerged as a promising paradigm for privacy-preserving edge intelligence.

Despite its advantages, practical wireless MFL is significantly more challenging than conventional unimodal FL. First, clients are modality-heterogeneous: each client may possess only a subset of modalities, and the available modalities can vary substantially across clients. Directly setting missing modalities to zero may introduce biased updates and degrade the final learning performance. Second, clients are resource-heterogeneous: their computing capability, battery budget, and uplink channel quality are often highly uneven [6]. Third, multimodal neural networks are generally more communication- and computation-intensive than unimodal networks, so requiring all scheduled clients to train and upload a full global model can easily violate latency constraints and aggravate the straggler effect. Although beamforming-aided uplink transmission with a multi-antenna edge server can improve wireless efficiency, scarce radio resources still make it necessary to jointly design learning and communication strategies for wireless MFL.

Motivated by these challenges, this paper studies a beamforming-aided wireless MFL system. The key idea is to allow each scheduled client to train a pruned unimodal submodel tailored to its communication and computing capabilities, while the server compensates for missing modalities using historical gradient information and aggregates updates in a storage-efficient manner. By integrating dynamic neural network pruning, modal compensation, client scheduling, and resource block (RB) allocation, the proposed framework aims to improve both learning efficiency and multimodal model quality under practical latency constraints.

A. Related Works

In wireless FL, extensive efforts have been devoted to addressing communication and computation bottlenecks in practical edge networks. Representative studies investigate convergence-time optimization via client selection [7], joint learning-and-communication design [8], energy-efficient wireless FL [9], wireless quantized FL [10], and robust FL under unreliable links [11]. Freshness-aware wireless FL has also been studied through age-of-update-based device selection in

Part of this work was presented in IEEE/CIC International Conference on Communications in China (ICCC), 2026 [1].

Chunfeng Xie, Zhixiong Chen and Arumugam Nallanathan are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. (emails: {c.xie, zhixiong.chen, a.nallanathan}@qmul.ac.uk).

W. Yi is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K. (email: wy23627@essex.ac.uk).

Chan-Tong Lam is with Faculty of Applied Sciences, Macao Polytechnic University, Macao (e-mail:ctlam@mpu.edu.mo).

a Stackelberg framework [12], age-weighted gradient aggregation under data heterogeneity [13], and joint device selection, sub-channel assignment, and resource allocation for AoI minimization [14]. [15] investigate intelligent reflecting surface (IRS)-aided FL through joint device scheduling, IRS phase-shift design, and communication–computation resource allocation, characterizing the learning latency–accuracy tradeoff under different multiple-access schemes. These studies motivate accounting for update freshness, while mFedDNP incorporates modality-region staleness into dynamic pruning and client–RB allocation. To further reduce latency and communication overhead, model-heterogeneous FL has also been studied through joint model pruning and device selection, as well as adaptive model pruning [16], [17]. In addition, recent multi-antenna FL works consider beamforming and device selection primarily in over-the-air aggregation settings [18], [19]. While these studies provide valuable insights into communication-efficient FL over wireless networks, they are largely developed for unimodal learning and do not explicitly address modality heterogeneity, multimodal fusion, or missing-modal updates.

On the MFL side, early and general frameworks mainly focus on decentralized fusion and aggregation across multiple sensing sources. For example, MMFed provides a unified multimodal FL framework for human activity recognition [20], FedMSplit enables correlation-adaptive federated multi-task learning across multimodal split networks [21], and FL-FD applies multimodal data fusion to federated fall detection [22]. Beyond task-specific designs, Harmony disentangles multimodal training into modality-wise federated learning and federated fusion to better handle heterogeneous nodes and latency imbalance [23], CreamFL supports heterogeneous client architectures and data modalities through contrastive representation ensemble and aggregation [24], and FedMultimodal provides a standardized benchmark for multimodal FL under corruptions such as missing modalities and missing labels [25]. Related directions also include cross-modal federated human activity recognition and unimodal-training/multimodal-prediction frameworks that aggregate knowledge across clients and modalities [26], [27].

More recent MFL studies have increasingly focused on modality heterogeneity, modal imbalance, and incomplete modalities. Hierarchical gradient blending has been proposed for non-IID multimodal FL [28], while MFedMC performs joint modality and client selection to reduce communication overhead [29]. Song *et al.* investigate modality-heterogeneous client drift [30], and recent balancing methods such as FedCMI and BMSFed address modality imbalance and modal bias through cross-modal knowledge transfer [31] and balanced modality selection [32]. Other works handle missing or incomplete modalities through imputation and distillation, such as AutoFed [33] and FedMEKT [34]; hybrid multimodal FL has also been studied by combining horizontal and vertical FL for multimodal IoT systems [35]. Nevertheless, most existing MFL studies abstract away the wireless training process and do not jointly optimize multi-antenna beamforming, client scheduling, adaptive neural network pruning, and RB allocation under round-latency constraints. This gap motivates the beamforming-aided wireless MFL framework developed in this paper. Related multimodal integrated sensing and communication (ISAC) studies coordinate visual–radar sensing and beamforming using heterogeneous mixture-of-experts controllers. Lyapunov-driven control targets tracking accuracy under computing-queue stability and long-term energy constraints [36], while reinforcement-learning-based control minimizes visual-processing and beam-recovery energy under sensing-freshness and link-reliability constraints [37]. mFedDNP addresses modality heterogeneity across federated clients through historical-gradient compensation and region-level pruning, jointly adapting client participation and RB allocation to update staleness and round-latency constraints. Movable-antenna ISAC jointly optimizes antenna positions and beamforming under sensing-SNR constraints [38]. This spatial flexibility could complement mFedDNP's receive combining, staleness-aware pruning, and client–RB allocation.

B. Motivations and Contributions

The above review reveals a clear gap between the wireless FL literature and the MFL literature. Existing wireless FL studies mainly focus on communication efficiency, latency reduction, and model adaptation for unimodal learning, whereas existing MFL studies mainly improve fusion, modality selection, and incomplete-modality learning at the algorithmic level. In practical wireless MFL, however, these issues are tightly coupled, since clients may have different available modalities, different uplink conditions, and different computing capabilities. As a result, forcing all scheduled clients to train and upload a full multimodal model is inefficient and may aggravate the straggler effect. These observations motivate a unified design that jointly adapts model size, client participation, and wireless resource allocation for multimodal federated learning. To this end, we propose a beamforming-aided wireless multimodal federated learning framework with dynamic neural network pruning, termed mFedDNP. The primary contributions of this study are outlined as follows.

- We develop a beamforming-aided wireless MFL framework for modality-heterogeneous clients over OFDMA uplinks. The proposed design integrates multimodal fusion, modal compensation, dynamic neural network pruning, and storage-efficient gradient-discrepancy maintenance, enabling efficient multimodal training under client heterogeneity.
- We characterize the impact of pruning on both transmitted parameters and local computation cost, and establish a latency model that jointly captures uplink transmission and local training delays. Based on this model, we formulate a joint optimization problem over client scheduling, neural network pruning, and RB allocation under per-round latency constraints.
- We analyze the convergence behavior of mFedDNP and derive a staleness-aware upper bound that explicitly reveals the effect of cumulative local unimodal gradient staleness on learning performance. This result further enables us to transform the original long-term loss minimization problem into a tractable round-level design objective.

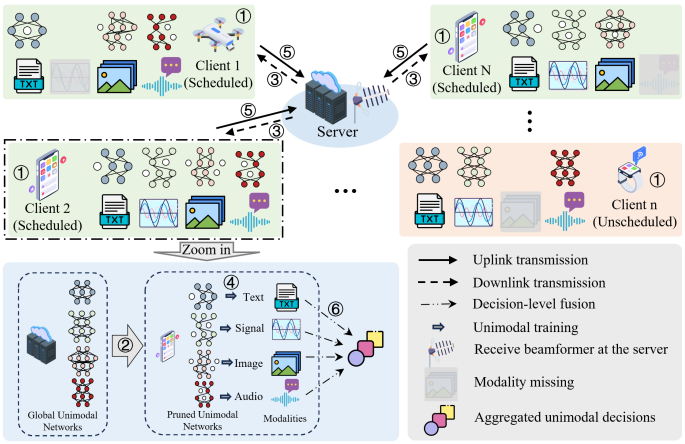


Fig. 1. Illustration of the mFedDNP over wireless networks: 1) Client Scheduling; 2) Neural Network Pruning; 3) Local Neural Network Downloading; 4) Local Neural Network Training with Modal Compensation; 5) Gradient Discrepancy Uploading; 6) Global Aggregation.

- We develop an efficient solution method by combining staleness-aware pruning with RB allocation and client scheduling, and validate the proposed framework on CREMA-D and UCI-HAR. Experimental results show that mFedDNP achieves better Macro-F1 and faster convergence than the benchmark schemes.

C. Organizations

The rest of this paper is structured as follows. Section II presents the beamforming-aided wireless MFL model and the proposed learning framework, including decision-level fusion, modal compensation, and dynamic neural network pruning. Section III provides the convergence analysis and introduces the staleness-aware design criterion. Section IV reformulates the optimization problem and develops the joint client scheduling, pruning, and RB allocation algorithm. Section V reports the experimental settings and numerical results. Section VI concludes the paper.

II. System Model and Learning Mechanism

We consider a wireless MFL system, illustrated in Fig. 1, in which the edge server is equipped with N_r receive antennas and each client has a single transmit antenna. The client set is denoted by $\mathcal{N} = \{1, 2, \cdots, N\}$. Let $\mathcal{M} = \{1, 2, \cdots, M\}$ represent the set of all modalities, and let $\mathcal{M}_n \subseteq \mathcal{M}$ denote the subset available at client n, where $|\mathcal{M}_n|$ is the number of modalities owned by client n. Accordingly, the local dataset of client n ($n \in \mathcal{N}$) is given by $\mathcal{D}_n = \{(\boldsymbol{x}_{n,m,s}, y_{n,s}) \mid s = \{1, 2, \cdots, D_n\}, m \in \mathcal{M}_n\}$, where D_n is the local data size, $\boldsymbol{x}_{n,m,s}$ is the feature vector of modality m for sample s, and $y_{n,s}$ is the corresponding label. The union of all local datasets is denoted by $\mathcal{D} = \bigcup_{n=1}^{N} \mathcal{D}_n$, and the total number of samples is $D = \sum_{n=1}^{N} D_n$. Let $\boldsymbol{w}$ denote the global multimodal model, written as $\boldsymbol{w} = [(\boldsymbol{v}_1)^\top, (\boldsymbol{v}_2)^\top, \cdots, (\boldsymbol{v}_M)^\top]^\top$, where $\boldsymbol{v}_m$ is the global unimodal submodel associated with modality m. To alleviate the straggler effect, each client is allowed to train a heterogeneous local network obtained by pruning the global unimodal submodels according to its communication and computation capabilities. Specifically, each unimodal submodel is divided into R disjoint regions indexed by $\mathcal{R} = \{1, 2, \cdots, R\}$, where each region corresponds to either a filter in a convolutional layer or a neuron in a fully connected layer. Let $\boldsymbol{v}_m^{(r)}$ denote the r-th region of $\boldsymbol{v}_m$. The output of $\boldsymbol{v}_m$ for modality m with input $\boldsymbol{x}_{n,m,s}$ is denoted by $\boldsymbol{v}_m \otimes \boldsymbol{x}_{n,m,s}$.

In conventional approaches, missing modalities are typically replaced by $\mathbf{0}$, which may introduce biased updates and degrade the learning performance. To address this issue, we design a modal compensation (MC) approach, detailed in Section II-A. For the analytical formulation, we use decision-level fusion in which the outputs of available modalities are averaged with equal weights. Accordingly, the multimodal loss of client n is given by

$$F_n(\boldsymbol{w}) = \frac{1}{D_n} \sum_{s \in \mathcal{D}_n} f\Big(y_{n,s}, \frac{1}{|\mathcal{M}_n|} \sum_{m \in \mathcal{M}_n} \boldsymbol{v}_m \otimes \boldsymbol{x}_{n,m,s}\Big), \quad (1)$$

where $f(\cdot)$ denotes the loss function, which is typically the cross-entropy between the *Softmax* output and the one-hot label in a classification task.

According to [29], [30], training unimodal submodels also contributes to multimodal performance, since the unimodal objective helps balance the convergence speeds of different modalities and accelerates the training of the multimodal model [31]. Therefore, we define the sum of unimodal losses at client n as

$$U_n(\boldsymbol{w}) = \sum\nolimits_{s \in \mathcal{D}_n} \sum\nolimits_{m=1}^{M} U_n(\boldsymbol{v}_{m,s}), \quad (2)$$

where $U_n(\boldsymbol{v}_{m,s}) = \frac{\xi_m}{D_n} \sum_{s \in \mathcal{D}_n} f(y_{n,s}, \boldsymbol{v}_{m,s} \otimes \boldsymbol{x}_{n,m,s})$ denotes the unimodal loss associated with modality m, and ξ_m is a preset modal weight. By combining the multimodal and unimodal objectives, the local loss of client n is

$$L_n(\boldsymbol{w}) = F_n(\boldsymbol{w}) + U_n(\boldsymbol{w}). \quad (3)$$

The global objective is defined as $L(\boldsymbol{w}) = \sum_{n=1}^{N} a_n L_n(\boldsymbol{w})$, where a_n denotes the aggregation weight of client n, satisfying $a_n \geq 0$ and $\sum_{n=1}^{N} a_n = 1$. The goal of the MFL system is to learn a shared global multimodal model $\boldsymbol{w}$ that minimizes the overall loss over the dataset $\mathcal{D}$, i.e., $\min_{\boldsymbol{w}} L(\boldsymbol{w})$.

A. Multimodal Federated Learning Mechanism

To enhance the communication and computation efficiency of wireless MFL, we propose the mFedDNP framework, which dynamically generates unimodal submodels for client-side training, as illustrated in Fig. 1. The training process proceeds over T global rounds, and each round $t \in \{0, 1, \cdots, T-1\}$ includes the following six steps:

1) **Client Scheduling**: The edge server selects a subset of clients to participate in the current round. Let $\alpha_{n,t} \in \{0, 1\}$ denote the scheduling indicator of client n in round t, where $\alpha_{n,t} = 1$ indicates that client n is selected and $\alpha_{n,t} = 0$ otherwise. For convenience, let $\mathcal{C}_t = \{n : \alpha_{n,t} = 1, \forall n \in \mathcal{N}\}$ denote the set of scheduled clients in round t.

2) **Neural Network Pruning**: After client scheduling, the edge server prunes the global unimodal networks to generate client-specific submodels according to the clients' computing and communication capabilities. Let $\boldsymbol{\beta}_{n,m,t} = \{\beta_{n,m,t}^{(r)} : r = 1, 2, \cdots, R\}$ denote the pruning mask for modality m of client

n in round t, where $\beta_{n,m,t}^{(r)}=1$ indicates that the r-th region of the global unimodal neural network is preserved in the local submodel of client n for modality m, and $\beta_{n,m,t}^{(r)}=0$ otherwise. Accordingly, the submodel of modality m at client n is written as $\boldsymbol{v}_{n,m,t}=\boldsymbol{v}_{m,t}\odot\boldsymbol{\beta}_{n,m,t}$, where $\odot$ denotes the element-wise product.

3) **Local Neural Network Downloading**: Each scheduled client downloads its pruned local model from the edge server.

4) **Local Neural Network Training with Modal Compensation**: For local updating, each scheduled client performs I-step mini-batch gradient descent with mini-batch $\mathcal{B}_{n,m,t}^{i}$ uniformly sampled from $\mathcal{D}_n$. The local multimodal gradient is given by $\nabla L_n(\boldsymbol{w}_t)=\nabla F_n(\boldsymbol{w}_t)+\nabla U_n(\boldsymbol{w}_t)$, where $\nabla L_n(\boldsymbol{w}_t)=[\nabla L_n^{\top}(\boldsymbol{v}_{1,t}),\nabla L_n^{\top}(\boldsymbol{v}_{2,t}),\cdots,\nabla L_n^{\top}(\boldsymbol{v}_{M,t})]^{\top}$ is composed of M unimodal subgradients. For $m\in\mathcal{M}_n$, the local unimodal subgradient of the r-th region ($\forall r\in\mathcal{R}$) is given by

$$\nabla\hat{L}_n(\boldsymbol{v}_{m,t,i+1}^{(r)})=\frac{\xi_m}{b}\sum\nolimits_{s\in\mathcal{B}_{n,m,t}^{i}}\nabla f(y_{n,s},\boldsymbol{v}_{m,t,i}^{(r)}\otimes\boldsymbol{x}_{n,m,s})+\frac{1}{b}\sum_{s\in\mathcal{B}_{n,m,t}^{i}}\frac{\partial f(y_{n,s},\frac{1}{|\mathcal{M}_n|}\sum_{m\in\mathcal{M}_n}\boldsymbol{v}_{m,t,i}^{(r)}\otimes\boldsymbol{x}_{n,m,s})}{\partial\boldsymbol{v}_{m,t,i}^{(r)}},\quad(4)$$

where $b=|\mathcal{B}_{n,m,t}^{i}|$ denotes the mini-batch size. To address the missing-modality issue, we adopt a modal compensation (MC) mechanism, which uses historical information instead of setting the corresponding unimodal loss to zero. In storage-efficient maintenance, client n retains its historical unimodal gradients $\{\boldsymbol{G}_{n,m,t}\}_{m=1}^{M}$, while the server maintains their aggregate $\{\hat{\boldsymbol{G}}_{m,t}\}_{m=1}^{M}$. For $m\notin\mathcal{M}_n$, MC supplies the corresponding update direction in (5) using the cached vector $\boldsymbol{G}_{n,m,t-1}^{(r)}$, retaining historical unimodal gradient information during modality absence. The i-th local update of the pruned multimodal model $\bar{\boldsymbol{w}}$ is then expressed as

$$\bar{\boldsymbol{w}}_{n,t,i+1}=\begin{bmatrix}\boldsymbol{v}_{n,1,t,i+1}^{(r)}\\ \boldsymbol{v}_{n,2,t,i+1}^{(r)}\\ \vdots\\ \boldsymbol{v}_{n,M,t,i+1}^{(r)}\end{bmatrix}=\begin{bmatrix}\boldsymbol{v}_{1,t,i}^{(r)}\\ \boldsymbol{v}_{2,t,i}^{(r)}\\ \vdots\\ \boldsymbol{v}_{M,t,i}^{(r)}\end{bmatrix}-\eta\begin{bmatrix}\nabla L_n(\boldsymbol{v}_{1,t,i}^{(r)})\\ \nabla L_n(\boldsymbol{v}_{2,t,i}^{(r)})\\ \vdots\\ \nabla L_n(\boldsymbol{v}_{M,t,i}^{(r)})\end{bmatrix},\quad(5)$$

where η is the learning rate.

5) **Gradient Discrepancy Uploading**: After local training, each scheduled client n ($n\in\mathcal{C}_t$) uploads the discrepancy between its current and previous unimodal subgradients, i.e., $\hat{\boldsymbol{G}}_{n,m,t}^{(r)}=\tilde{\boldsymbol{g}}_{n,m,t}^{(r)}-\boldsymbol{G}_{n,m,t-1}^{(r)}$, to the edge server, where $\tilde{\boldsymbol{g}}_{n,m,t}^{(r)}$ is given by $\tilde{\boldsymbol{g}}_{n,m,t}^{(r)}=\sum_{i=0}^{I-1}\nabla\hat{L}_n(\boldsymbol{v}_{m,t,i}^{(r)})$.

6) **Global Aggregation with Storage-Efficient Maintenance**: After receiving the uploaded gradient discrepancies, the edge server updates the aggregated unimodal gradient as $\hat{\boldsymbol{G}}_{m,t}^{(r)}=\hat{\boldsymbol{G}}_{m,t-1}^{(r)}+\frac{1}{N}\sum_{n=1}^{N}\alpha_{n,t}\beta_{n,m,t}^{(r)}\hat{\boldsymbol{G}}_{n,m,t}^{(r)}$, and updates region r of the global unimodal submodel by $\boldsymbol{v}_{m,t+1}^{(r)}=\boldsymbol{v}_{m,t}^{(r)}-\eta\hat{\boldsymbol{G}}_{m,t}^{(r)}$. Accordingly, the global loss, defined as a weighted sum of local losses, is given by

$$\begin{aligned}L(\bar{\boldsymbol{w}}_t)&=\sum\nolimits_{n\in\mathcal{C}_t}a_nL_n(\bar{\boldsymbol{w}}_t)\\&=\sum\nolimits_{n\in\mathcal{C}_t}a_nF_n(\bar{\boldsymbol{w}}_t)+\sum\nolimits_{n\in\mathcal{C}_t}a_nU_n(\bar{\boldsymbol{w}}_t)\\&=F(\bar{\boldsymbol{w}}_t)+U(\bar{\boldsymbol{w}}_t),\end{aligned}\quad(6)$$

where $a_n=\frac{D_n}{\sum_{n\in\mathcal{C}_t}D_n}$. After T communication rounds, the loss of the pruned global multimodal neural network $\bar{\boldsymbol{w}}_T$ is expected to be minimized. Therefore, the objective of MFL is

$$\min_{\bar{\boldsymbol{w}}_T}L(\bar{\boldsymbol{w}}_T)=F(\bar{\boldsymbol{w}}_T)+U(\bar{\boldsymbol{w}}_T).\quad(7)$$

In the following theorem, we establish the efficiency of the proposed maintenance mechanism.

Theorem 1. *The proposed storage-efficient maintenance mechanism distributes the memory burden between the edge server and the clients as the model size and the number of participating clients increase. Specifically, maintaining the aggregated unimodal gradient $\hat{\boldsymbol{G}}_{m,t}^{(r)}$ yields the same global unimodal update as maintaining the full client-specific gradient array $\{\boldsymbol{G}_{n,m,t}^{(r)}\}_{n=1}^{N}$ at the edge server.*

Proof: Please see Appendix A. ∎

MC reuses the latest available unimodal gradient information with unit compensation weight, while staleness enters the pruning and scheduling criterion. Age-dependent expiration, attenuation, and confidence-weighted compensation offer future extensions for prolonged modality absence.

B. Learning Latency Model

In this subsection, we first relate the pruning ratio to the communication and computation costs of the pruned models, and then establish the per-round latency model.

1) Pruned Network Size Analysis: In each global round, client n trains a pruned unimodal submodel to match its communication and computing capabilities. For modality m of client n in round t, the pruning ratio under mask $\boldsymbol{\beta}_{n,m,t}$ is defined as

$$q_{n,m,t}=1-\frac{1}{R}\sum\nolimits_{r=1}^{R}\beta_{n,m,t}^{(r)}.\quad(8)$$

The pruning ratio $q_{n,m,t}$ is the fraction of removed regions, so $R(1-q_{n,m,t})$ regions are retained. The same pruning ratio is applied across the convolutional and fully connected layers of each unimodal submodel. To connect pruning to learning cost, let $\mathcal{Q}_m^{\mathrm{comm}}$ denote the total number of parameters in the global unimodal model for modality m, and let $\mathcal{Q}_m^{\mathrm{comp}}$ denote the number of floating-point operations (FLOPs) required to process one data sample. In addition, let $Q_{m,j}^{\mathrm{conv}}$ and $Q_{m,j}^{\mathrm{FC}}$ denote the sizes of the j-th convolutional and FC layers, respectively.

For the j-th convolutional layer, the number of parameters is $\mathcal{Q}_{m,j}^{comm}=Q_{m,j}^{conv}\times(\kappa^2\times Q_{m,j-1}^{conv}+1)$, where κ^2 is the kernel size. After pruning with ratio $q_{n,m,t}$, the parameter count becomes $\mathcal{Q}_{n,m,j}^{comm}=(1-q_{n,m,t})^2\times Q_{m,j}^{conv}\times\kappa^2\times Q_{m,j-1}^{conv}+(1-q_{n,m,t})\times Q_{m,j}^{conv}\approx(1-q_{n,m,t})^2\mathcal{Q}_{m,j}^{comm}$. Similarly, for the j-th FC layer, $\mathcal{Q}_{m,j}^{comm}=(Q_{m,j-1}^{FC}+1)\times Q_{m,j}^{FC}$, and the corresponding pruned parameter count is $\mathcal{Q}_{n,m,j}^{comm}=(1-q_{n,m,t})^2\times Q_{m,j-1}^{FC}\times Q_{m,j}^{FC}+(1-q_{n,m,t})\times Q_{m,j}^{FC}\approx(1-q_{n,m,t})^2\mathcal{Q}_{m,j}^{comm}$. Since the number of bias parameters is much smaller than the number of weight parameters, the total

number of parameters in the pruned unimodal submodel can be approximated as $\mathcal{Q}_{n,m}^{comm} = (1-q_{n,m,t})^2 \mathcal{Q}_m^{comm}$.

The same quadratic scaling also applies to the computational cost. For the j-th convolutional layer, the number of FLOPs is $\mathcal{Q}_{m,j}^{comp} = 2\kappa^2 \mathrm{wh} Q_{m,j}^{conv} \times Q_{m,j-1}^{conv}$, where w and h denote the width and height of the input feature map, respectively. After pruning, $\mathcal{Q}_{n,m,j}^{comp} = 2\kappa^2 \mathrm{wh}(1-q_{n,m,t})^2 \times Q_{m,j}^{conv} \times Q_{m,j-1}^{conv} = (1-q_{n,m,t})^2 \mathcal{Q}_{m,j}^{comp}$. For the j-th FC layer, the pruned FLOP count is $\mathcal{Q}_{n,m,j}^{comp} = 2(1-q_{n,m,t})^2 \times Q_{m,j-1}^{FC} \times Q_{m,j}^{FC} = (1-q_{n,m,t})^2 \mathcal{Q}_{m,j}^{comp}$. Therefore, the total computational cost of the pruned unimodal submodel is $\mathcal{Q}_{n,m}^{comp} = (1-q_{n,m,t})^2 \mathcal{Q}_m^{comp}$.

2) Beamforming-aided Communication Model: For uplink transmission, we consider an OFDMA system with K resource blocks (RBs). Let $\mathcal{K} = \{1, 2, \cdots, K\}$ denote the set of available RBs, and let $\boldsymbol{\theta}_{n,t} = (\theta_{n,t}^{(1)}, \theta_{n,t}^{(2)}, \cdots, \theta_{n,t}^{(K)})$ denote the RB allocation vector of client n in round t, where $\theta_{n,t}^{(k)} \in \{0,1\}$. Here, $\theta_{n,t}^{(k)} = 1$ means that RB k is assigned to client n, and $\theta_{n,t}^{(k)} = 0$ otherwise. Each client can occupy at most one RB, i.e., $\sum_{k=1}^{K} \theta_{n,t}^{(k)} \leq 1$, and each RB can be assigned to at most one client, i.e., $\sum_{n=1}^{N} \theta_{n,t}^{(k)} \leq 1$. The RB allocation for all clients in round t is denoted by $\boldsymbol{\Theta}_t = (\boldsymbol{\theta}_{1,t}, \boldsymbol{\theta}_{2,t}, \cdots, \boldsymbol{\theta}_{N,t})$. We also consider inter-cell interference δ_k caused by users in other service areas that reuse the same RB [7]. Let $\boldsymbol{f}_{n,t}^{(k)} \in \mathbb{C}^{N_r \times 1}$ denote the receive beamformer at the server for client n on RB k. The uplink rate of client n is then given by

$$\mathrm{r}_{n,t}(\boldsymbol{\theta}_{n,t}, \boldsymbol{f}_{n,t}) = \sum_{k=1}^{K} \theta_{n,t}^{(k)} B \log_2\left(1 + \frac{p_n |\boldsymbol{f}_{n,t}^{(k)H} \boldsymbol{h}_{n,t}^{(k)}|^2}{\|\boldsymbol{f}_{n,t}^{(k)}\|^2 (B\nu_0 + \delta_k)}\right), \tag{9}$$

where B is the bandwidth of one RB, p_n is the transmit power of client n, $\boldsymbol{h}_{n,t}^{(k)} \in \mathbb{C}^{N_r \times 1}$ is the uplink channel from client n to the edge server on RB k, and ν_0 is the noise power spectral density. Let ϵ denote the number of quantization bits for each uploaded parameter. Then, the communication latency of client n is

$$\mathrm{T}_{n,t}^{\mathrm{cm}} = \frac{\epsilon \sum_{m=1}^{M} \mathcal{Q}_{n,m}^{comm}}{\mathrm{r}_{n,t}(\boldsymbol{\theta}_{n,t}, \boldsymbol{f}_{n,t})}. \tag{10}$$

3) Computing Latency Model: For local computation, we assume that each client performs training using a CPU, where each CPU cycle can process q_n FLOPs. Let f_n denote the CPU frequency of client n. Then, the computation latency is

$$\mathrm{T}_{n,t}^{\mathrm{cp}} = \frac{Ib \sum_{m=1}^{M} \mathcal{Q}_{n,m}^{comp}}{q_n f_n}. \tag{11}$$

Finally, we note that the latencies of neural network updating, network downloading, network fusion, and network pruning are negligible compared with the communication and local training latencies. In addition, the edge server is assumed to have much stronger computing and transmission capabilities than the edge clients [10].

C. Problem Formulation

In this subsection, we formulate an optimization problem to minimize the global loss after T communication rounds, thereby improving the learning performance of the proposed mFedDNP framework. Since each RB can be assigned to at most one scheduled client, there is no intra-cell multiuser interference among participating clients. Therefore, the edge server adopts maximum ratio combining (MRC) for each scheduled uplink transmission, i.e., the receive beamformer is aligned with the corresponding channel vector. We jointly optimize the client scheduling, neural network pruning, and RB allocation policies under wireless resource and latency constraints, and formulate the problem as

$$\mathcal{P}: \quad \min_{\{\mathcal{C}_t, \boldsymbol{\beta}_t, \boldsymbol{\Theta}_t\}_{t=0}^{T-1}} \mathbb{E}[L(\bar{\boldsymbol{w}}_T)] \tag{12}$$

$$\text{s. t.} \quad \mathrm{T}_{n,t}^{\mathrm{cm}} + \mathrm{T}_{n,t}^{\mathrm{cp}} \leq \mathrm{T}_{\max}, \forall n \in \mathcal{C}_t, \forall t, \tag{12a}$$

$$q_{n,m,t} = 1 - \frac{1}{R} \sum_{r=1}^{R} \beta_{n,m,t}^{(r)}, \forall n \in \mathcal{N}, \forall m \in \mathcal{M}, \forall t, \tag{12b}$$

$$0 \leq q_{n,m,t} \leq 1, \forall n \in \mathcal{N}, \forall t, \tag{12c}$$

$$\sum_{k=1}^{K} \theta_{n,t}^{(k)} \leq 1, \forall n \in \mathcal{N}, \forall t, \tag{12d}$$

$$\sum_{n=1}^{N} \theta_{n,t}^{(k)} \leq 1, \forall t, \tag{12e}$$

$$\theta_{n,t}^{(k)} \in \{0,1\}, \forall n \in \mathcal{N}, \forall t, \tag{12f}$$

$$\beta_{n,m,t}^{(r)} \in \{0,1\}, \forall n \in \mathcal{N}, \forall m \in \mathcal{M}, \forall t, \tag{12g}$$

$$\alpha_{n,t} \in \{0,1\}, \forall n \in \mathcal{N}, \forall t, \tag{12h}$$

where (12a) limits the training and uplink latency of scheduled clients, with $\alpha_{n,t} = \sum_{k=1}^{K} \theta_{n,t}^{(k)}$. The latency expressions in (10) and (11) are evaluated for $n \in \mathcal{C}_t$; unscheduled clients perform no local training or uplink transmission in that round. (12b) indicates the connection between the pruning ratio and pruning policy for clients on all modalities. (12c) restricts the value of pruning ratio. (12d), (12e) and (12f) correspond to the constraints on the RBs allocation policy. (12g) is the constraint on decision of pruning region. (12h) indicates the client scheduling decision in each round.

It is clear that $\mathcal{P}$ is difficult to solve directly for two reasons. First, it involves multi-dimensional discrete variables and is therefore a challenging integer programming problem. Second, there is no explicit expression for $\mathbb{E}[L(\bar{\boldsymbol{w}}_T)]$ with respect to $\mathcal{C}_t$, $\boldsymbol{\beta}_t$, and $\boldsymbol{\Theta}_t$, since the evolution of the learning model over multiple rounds is highly complex. Therefore, in the next section, we derive an upper bound on $\mathbb{E}[L(\bar{\boldsymbol{w}}_T)]$ and minimize that bound instead to guide the global loss minimization [7]–[9], [17].

III. Convergence Analysis

In this section, we analyze the convergence behavior of mFedDNP to reveal how client scheduling, neural network pruning, and RB allocation affect the learning performance. Motivated by the derived convergence bound, we further transform problem $\mathcal{P}$ into a more tractable form with a new design objective for wireless MFL.

For $m \in \mathcal{M}_n$, $\nabla L_n(\boldsymbol{v}_{m,t,i})$ denotes the modality-m block of $\nabla L_n(\boldsymbol{w})$, evaluated at the i-th local iterate using the full local dataset $\mathcal{D}_n$. Let $L(\boldsymbol{w}^*)$ denote the loss value of the optimal global multimodal model $\boldsymbol{w}^*$. Under the MC mechanism, the global model is updated using the latest available unimodal gradients, which helps alleviate the negative effect of partial

participation. To capture the freshness of local gradients, we define the staleness of the local unimodal subgradient of client n on modality m at region r as

$$\tau_{n,m,t}^{(r)} = \begin{cases} \tau_{n,m,t-1}^{(r)} + 1, & \text{if } \alpha_{n,t}\beta_{n,m,t}^{(r)} = 0. \\ 0, & \text{else.} \end{cases} \tag{13}$$

We next introduce the following standard assumptions on the local loss functions.

Assumption 1. The local loss function is L-smooth. That is, for all $\boldsymbol{w}$ and $\boldsymbol{w}'$, we have $\|\nabla L_n(\boldsymbol{w}) - \nabla L_n(\boldsymbol{w}')\| \leq L\|\boldsymbol{w} - \boldsymbol{w}'\|$.

Assumption 2. The local loss function is ρ-strongly convex. That is, for all $\boldsymbol{w}$ and $\boldsymbol{w}'$, we have $L_n(\boldsymbol{w}') \geq L_n(\boldsymbol{w}) + \langle \nabla L_n(\boldsymbol{w}), \boldsymbol{w}' - \boldsymbol{w}\rangle + \frac{\rho}{2}\|\boldsymbol{w}' - \boldsymbol{w}\|^2$.

Assumption 3. The stochastic gradient $\nabla \hat{L}_n(\boldsymbol{v}_{m,t})$ is an unbiased estimator of the full gradient $\nabla L_n(\boldsymbol{v}_{m,t})$, i.e., $\mathbb{E}[\nabla \hat{L}_n(\boldsymbol{v}_{m,t})] = \nabla L_n(\boldsymbol{v}_{m,t})$, and its variance is bounded as $\mathbb{E}\|\nabla \hat{L}_n(\boldsymbol{v}_{m,t}) - \nabla L_n(\boldsymbol{v}_{m,t})\|^2 \leq \sigma^2$.

Assumption 4. The squared norm of the local unimodal gradient is uniformly bounded, i.e., $\|\nabla L_n(\boldsymbol{v}_{m,t})\|^2 \leq \Phi^2$.

Assumptions 1-4 are widely adopted in the FL and MFL literature for convergence analysis, e.g., [8], [9], [28], [35]. Assumptions 1 and 2 characterize the loss function, while Assumptions 3 and 4 describe the stochastic training process of the unimodal models. Based on these assumptions, we first present the following auxiliary lemma.

Lemma 1. *Let Assumptions* 1, 3, *and* 4 *hold. If the learning rate satisfies* $\eta \leq \frac{1}{2IL}$*, then the averaged difference between the global neural networks at two rounds* t *and* t' *with* $t \geq t'$ *is bounded as*

$$\mathbb{E}\|\boldsymbol{v}_{m,t}^{(r)} - \boldsymbol{v}_{m,t'}^{(r)}\|^2 \leq 3\eta^2(t-t')^2[(2R(I-1)+I^2)\Phi^2 + ((I-1)R+I^2)\sigma^2)]. \tag{14}$$

Proof: Please see Appendix B. ∎

Based on Lemma 1, we next derive the one-round convergence bound of mFedDNP.

Theorem 2. *Under Assumptions* 1, 3, *and* 4, *and for a learning rate satisfying* $\eta \leq \frac{1}{2IL}$*, the one-round convergence bound can be written as*

$$\mathbb{E}[L(\boldsymbol{w}_{t+1}) - L(\boldsymbol{w}_t)] \leq (-\frac{1}{2}I\eta + I^2L\eta^2)\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 + \gamma_1 + \frac{\gamma_2}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{r=1}^{R}(1-\alpha_{n,t}\beta_{n,m,t}^{(r)})(\tau_{n,m,t-1}^{(r)}+1)^2, \tag{15}$$

where $\gamma_1 = \frac{3}{4}I\eta\sigma^2 + 8IL^2RM\eta(I-1)(2I\eta^2\Phi^2 + \eta^2\sigma^2)$, $\gamma_2 = \frac{15}{2}IL^2\eta^3[(2R(I-1)+I^2)\Phi^2 + ((I-1)R+I^2)\sigma^2)]$.

Proof: Please see Appendix C. ∎

According to Theorem 2, the term $\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{r=1}^{R}(1-\alpha_{n,t}\beta_{n,m,t}^{(r)})(\tau_{n,m,t-1}^{(r)}+1)^2$ represents the sum of squared staleness values of the local subgradients, which is a key factor that negatively affects the convergence rate. Therefore, properly designing the client scheduling, neural network pruning, and RB allocation policies can tighten the convergence bound and improve the learning performance.

Theorem 2 provides a one-round bound under Assumptions 1, 3, and 4, whose staleness term motivates the policy evaluated with the MLP architecture in Section V-A. Corollary 1 additionally invokes strong convexity to characterize the multi-round optimality gap.

Corollary 1. *Under Assumptions* 1-4*, the expected gap between the global loss after* T *rounds and the optimal value is bounded as*

$$\begin{aligned} &\mathbb{E}[L(\boldsymbol{w}_T) - L(\boldsymbol{w}^*)] \\ &\leq (1 - I\eta\rho + 2I^2L\eta^2\rho)^T\mathbb{E}[L(\boldsymbol{w}_0) - L(\boldsymbol{w}^*)] \\ &+ \frac{1-(1-I\eta\rho+2I^2L\eta^2\rho)^T}{I\eta\rho - 2I^2L\eta^2\rho}\gamma_1 \\ &+ \frac{\gamma_2}{N}\sum_{t=1}^{T-1}(1-I\eta\rho+2I^2L\eta^2\rho)^{T-1-t}\times \\ &\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{r=1}^{R}(1-\alpha_{n,t}\beta_{n,m,t}^{(r)})(\tau_{n,m,t-1}^{(r)}+1)^2. \end{aligned} \tag{16}$$

Proof: Please see Appendix D. ∎

Corollary 1 indicates that the expected optimality gap after T rounds can be decomposed into three components: i) the initial gap between $L(\boldsymbol{w}_0)$ and $L(\boldsymbol{w}^*)$; ii) a constant term determined by the system hyperparameters, the number of local epochs I, and the stochastic noise; and iii) the cumulative staleness of local subgradients over the T rounds. The first two components are fixed once the learning setup is specified, whereas the third depends directly on the client scheduling, neural network pruning, and RB allocation policies. Therefore, reducing the last term is essential for improving the overall learning performance. This insight motivates the following remark.

Remark 1. Following several existing works, e.g., [8], [9], [11], we assume that the available wireless resources and the set of participating clients in problem $\mathcal{P}$ are independent across communication rounds. Under this assumption, the convergence bound in (16) can be tightened by minimizing the last term on its right-hand side. Motivated by Theorem 2 and Corollary 1, we therefore transform the original long-term optimization into a round-level objective, $\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{r=1}^{R}(1-\alpha_{n,t}\beta_{n,m,t}^{(r)})(\tau_{n,m,t-1}^{(r)}+1)^2$, which reduces the upper bound on $\mathbb{E}[L(\boldsymbol{w}_{t+1}) - L(\boldsymbol{w}_t)]$ at each round and, in turn, helps decrease the overall T-round convergence bound in (16).

IV. Problem Transformation and Solution

In this section, we develop an efficient algorithm for joint client scheduling, neural network pruning, and RB allocation to solve problem $\mathcal{P}$. To this end, we first transform problem $\mathcal{P}$ into a more tractable form based on the convergence analysis in Section III.

A. Problem Transformation

The results in Theorem 2 and Corollary 1 reveal how the wireless design affects the learning performance of mFedDNP. According to Remark 1, problem $\mathcal{P}$ can be transformed into minimizing $\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{r=1}^{R}(1-\alpha_{n,t}\beta_{n,m,t}^{(r)})(\tau_{n,m,t-1}^{(r)}+1)^2$ in each round through client scheduling, RB allocation, and dynamic neural network pruning. Moreover, since

$\alpha_{n,t} = \sum_{k=1}^{K} \theta_{n,t}^{(k)} \in \{0,1\}$, the client scheduling decision is fully determined once the RB allocation is given. Therefore, problem $\mathcal{P}$ can be reformulated as a round-level problem that minimizes the sum of squared staleness values of local subgradients:

$$\widehat{\mathcal{P}}: \quad \min_{\boldsymbol{\beta}_t, \boldsymbol{\Theta}_t} \ \mathcal{W}(\boldsymbol{\beta}_t, \boldsymbol{\Theta}_t) \quad \text{s. t.} \ \ (12a) - (12g). \tag{17}$$

where $\mathcal{W}(\boldsymbol{\beta}_t, \boldsymbol{\Theta}_t) = \sum_{n=1}^{N} \sum_{m=1}^{M} \sum_{r=1}^{R} (1 - \sum_{k=1}^{K} \theta_{n,t}^{(k)} \beta_{n,m,t}^{(r)})(\tau_{n,m,t-1}^{(r)} + 1)^2$. Problem $\widehat{\mathcal{P}}$ is still a challenging integer programming problem. In the following, we first derive the dynamic neural network pruning policy for each client under a given RB allocation decision, and then transform problem $\widehat{\mathcal{P}}$ into an equivalent linear programming problem that can be solved efficiently.

B. Dynamic Neural Network Pruning

For a given RB allocation policy $\boldsymbol{\Theta}_t$, the pruning decisions of different clients are independent in the objective function. Therefore, the pruning policy of each scheduled client can be optimized separately. Based on this observation, for each scheduled client n ($n \in \mathcal{C}_t$), problem $\widehat{\mathcal{P}}$ can be decomposed into the following pruning subproblem:

$$\widehat{\mathcal{P}}_1: \quad \max_{\boldsymbol{\beta}_{n,t}} \ \sum_{k=1}^{K} \theta_{n,t}^{(k)} \sum_{m=1}^{M} \sum_{r=1}^{R} \beta_{n,m,t}^{(r)} (\tau_{n,m,t-1}^{(r)} + 1)^2 \quad \text{s. t.} \ \ (12a), (12b), (12c), (12g). \tag{18}$$

For a fixed RB assignment, let $\bar{q}_{n,m,t} = \sum_{r=1}^{R} \beta_{n,m,t}^{(r)}$ denote the preserved-region count of modality m, and define $c_{n,m,t} = \epsilon \mathcal{Q}_m^{\text{comm}} / \mathrm{r}_{n,t} + Ib \mathcal{Q}_m^{\text{comp}} / (q_n f_n)$, where $\mathrm{r}_{n,t} > 0$ is the assigned uplink rate. Under the quadratic workload model in Section II-B, the latency constraint becomes

$$\sum_{m=1}^{M} c_{n,m,t} \left(\frac{\bar{q}_{n,m,t}}{R} \right)^2 \le \mathrm{T}_{\max}, \qquad \bar{q}_{n,m,t} \in \{0, 1, \ldots, R\}, \qquad \forall m \in \mathcal{M}. \tag{19}$$

We enumerate the admissible modality-wise count vectors under the width-scaling rule. For each vector satisfying (19), the highest-staleness regions are retained within its layer-wise quotas. The vector maximizing (18) and its corresponding masks are selected.

Remark 2. For fixed modality-wise counts and their associated layer-wise quotas, retaining the highest-staleness regions within each layer maximizes the objective in (18) without changing the modeled latency.

The staleness score prioritizes refreshing regions with longer update gaps according to the delay-dependent term in Section III. The common smoothness and gradient bounds yield a shared freshness criterion across modalities with different gradient scales and learning dynamics. Modality-specific communication and computation costs in (19) determine the feasible retained counts. Similar to several existing works, e.g., [16], [17], [39], we adopt width scaling to prune the global model, where a certain number of filters or neurons are removed from each convolutional or FC layer to generate a client-specific submodel. For each unimodal submodel, the same pruning ratio is used across its convolutional and fully connected layers. For each layer, the filters or neurons are first sorted in descending order according to their gradient staleness, and then the top regions are preserved according to (19), while the remaining ones are removed. Sorting can be implemented by standard algorithms such as Quicksort or Introsort, both with time complexity $\mathcal{O}(S_j \log S_j)$, where S_j denotes the number of filters or neurons in the j-th layer. The layer-wise sorting has complexity $\mathcal{O}(S \log(\max_{j \in \mathcal{J}} S_j))$.

Algorithm 1 Dynamic Client Scheduling, Neural Network Pruning, and RB Allocation Algorithm

Input: Region staleness $\{\tau_{n,m,t-1}^{(r)}\}$, modality-specific model costs, client computing resources, candidate uplink rates $\{\mathrm{r}_{n,t}^{(k)}\}$, and deadline $\mathrm{T}_{\max}$.
Initialize all pair-feasibility flags to false, $\boldsymbol{\Theta}_t^* = \mathbf{0}$, and $\boldsymbol{\beta}_t^* = \mathbf{0}$.
for each client $n \in \mathcal{N}$ **do**
  Sort the regions within each layer in descending order of gradient staleness.
  Precompute the retained-score sums for each admissible modality-wise count under the width-scaling rule.
  for each RB $k \in \mathcal{K}$ with $\mathrm{r}_{n,t}^{(k)} > 0$ **do**
    Enumerate admissible modality-wise count vectors and retain those satisfying (19) at rate $\mathrm{r}_{n,t}^{(k)}$.
    if at least one feasible count vector exists **then**
      Select the feasible count vector maximizing (18), retain the highest-staleness regions within its layer-wise quotas, and store $\{\boldsymbol{\beta}_{n,m,t}^{*,k}\}_{m=1}^{M}$.
      Mark (n,k) feasible and compute its weight $\Lambda_{n,k}$ as the sum of the retained squared-staleness scores.
Solve (20) by maximum-weight bipartite matching over feasible client–RB pairs. Set $\theta_{n,t}^{*,(k)} = 1$ for matched pairs and zero otherwise.
Compute $\alpha_{n,t}^* = \sum_{k=1}^{K} \theta_{n,t}^{*,(k)}$ and $\mathcal{C}_t^* = \{n \in \mathcal{N} : \alpha_{n,t}^* = 1\}$.
for each scheduled client $n \in \mathcal{C}_t^*$ **do**
  Let k be its assigned RB and set $\boldsymbol{\beta}_{n,m,t}^* = \boldsymbol{\beta}_{n,m,t}^{*,k}$ for every modality m.
return $\mathcal{C}_t^*$, $\boldsymbol{\beta}_t^*$, and $\boldsymbol{\Theta}_t^*$.

C. Client Scheduling and Resource Block Allocation

Let $\mathcal{F}_t$ contain the real client–RB pairs with a positive uplink rate and an admissible pruning configuration satisfying (19). For $(n,k) \in \mathcal{F}_t$, let $\boldsymbol{\beta}_{n,m,t}^{*,k}$ denote the corresponding optimal pruning mask. For $(n,k) \notin \mathcal{F}_t$, set $\theta_{n,t}^{(k)} = 0$ and use $\boldsymbol{\beta}_{n,m,t}^{*,k} = \mathbf{0}$ only for score bookkeeping. By substituting the optimal pruning policies of all clients into problem $\widehat{\mathcal{P}}$, we obtain the following equivalent RB allocation problem:

$$\widehat{\mathcal{P}}_2: \quad \max_{\boldsymbol{\Theta}_t} \ \sum_{n=1}^{N} \sum_{k=1}^{K} \theta_{n,t}^{(k)} \sum_{m=1}^{M} \sum_{r=1}^{R} \beta_{n,m,t}^{*,k,(r)} (\tau_{n,m,t-1}^{(r)} + 1)^2 \quad \text{s. t.} \ \ (12d), (12e), (12f). \tag{20}$$

Problem $\widehat{\mathcal{P}}_2$ is still an integer programming problem. To solve it efficiently, we transform it into a maximum-weight perfect bipartite matching problem [40], which can be solved in polynomial time. In a bipartite matching problem, the goal is to select a set of non-overlapping edges with the maximum total weight.

To this end, we construct a complete and balanced bipartite graph $\Omega = (\mathcal{S}, \mathcal{U})$, where the vertex set is $\mathcal{S} = \mathcal{N} \cup \overline{\mathcal{K}}$ and

$\mathcal{U}$ is the set of edges connecting vertices in $\mathcal{N}$ and $\overline{\mathcal{K}}$. Each vertex $n \in \mathcal{N}$ represents client n. The set $\overline{\mathcal{K}} = \mathcal{K} \cup \mathcal{K}_u$ extends the original RB set $\mathcal{K}$ by adding a virtual RB set $\mathcal{K}_u$, such that $|\overline{\mathcal{K}}| = |\mathcal{N}|$. This construction makes the bipartite graph balanced. The edge weight is defined as

$$\Lambda_{n,k} = \begin{cases} \sum_{m=1}^{M} \sum_{r=1}^{R} \beta_{n,m,t}^{*,k,(r)} (\tau_{n,m,t-1}^{(r)} + 1)^2, & (n,k) \in \mathcal{F}_t, \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

In this paper, we focus on the case $|\mathcal{N}| \geq |\mathcal{K}|$. When $|\mathcal{K}| > |\mathcal{N}|$, one can similarly introduce a virtual client set $\mathcal{N}_u$ such that $|\mathcal{N}| + |\mathcal{N}_u| = |\mathcal{K}|$ and construct an equivalent balanced graph. Based on the above graph, problem $\widehat{\mathcal{P}}_2$ can be transformed into the following maximum-weight perfect bipartite matching problem. Let $\omega_{n,k} \in \{0,1\}$ indicate selection of an edge in the complete auxiliary graph. Zero-weight entries outside $\mathcal{F}_t$ are completion edges representing no physical transmission. Recover $\theta_{n,t}^{(k)} = \omega_{n,k}$ for $(n,k) \in \mathcal{F}_t$ and zero otherwise. Under this recovery rule, the nonnegative feasible-edge weights ensure equality of the optimal physical-allocation and perfect-matching values. For convenience, let $\boldsymbol{\omega}_n = \{\omega_{n,1}, \cdots, \omega_{n,|\overline{\mathcal{K}}|}\}$ denote the RB assignment vector of client n. The problem is then written as

$$\mathcal{P}_3: \quad \max_{\{\boldsymbol{\omega}_n\}_{n=1}^N} \quad \sum\nolimits_{n=1}^{N} \sum\nolimits_{k=1}^{|\overline{\mathcal{K}}|} \omega_{n,k} \Lambda_{n,k} \tag{22}$$

$$\text{s. t.} \quad \sum\nolimits_{k=1}^{|\overline{\mathcal{K}}|} \omega_{n,k} = 1, \forall n \in \mathcal{N}, \tag{22a}$$

$$\sum\nolimits_{n=1}^{N} \omega_{n,k} = 1, \forall k \in \overline{\mathcal{K}}, \tag{22b}$$

$$\omega_{n,k} \in \{0,1\}, \forall n \in \mathcal{N}, \forall k \in \overline{\mathcal{K}}. \tag{22c}$$

Any feasible solution to $\mathcal{P}_3$ corresponds to a perfect matching of Ω. However, $\mathcal{P}_3$ is still a linear integer programming problem. To further simplify it, we relax the integrality constraint (22c) and obtain the following linear programming problem:

$$\widehat{\mathcal{P}}_3: \quad \max_{\{\boldsymbol{\omega}_n\}_{n=1}^N} \quad \sum\nolimits_{n=1}^{N} \sum\nolimits_{k=1}^{|\overline{\mathcal{K}}|} \omega_{n,k} \Lambda_{n,k} \tag{23}$$

$$\text{s. t.} \quad (22a), (22b)$$

$$0 \leq \omega_{n,k} \leq 1, \forall n \in \mathcal{N}, \forall k \in \overline{\mathcal{K}}. \tag{23a}$$

Problem $\widehat{\mathcal{P}}_3$ is the linear programming relaxation of $\mathcal{P}_3$. It can be solved by the current matrix multiplication time algorithm [41], with time complexity $\mathcal{O}((N^{2+\frac{1}{6}})^2)$, since it contains N^2 variables, i.e., $\omega_{n,k}$ for $n \in \mathcal{N}$ and $k \in \overline{\mathcal{K}}$. The constraint matrix associated with (22a) and (22b) is the vertex–edge incidence matrix of a bipartite graph. Multiplying the rows of one vertex set by -1 gives a directed incidence matrix, establishing total unimodularity [40]. Thus, $\widehat{\mathcal{P}}_3$ admits an integral optimal extreme point, which solves $\mathcal{P}_3$. The physical allocation is then obtained using the stated recovery rule. In summary, we first transform problem $\widehat{\mathcal{P}}$ into the equivalent maximum-weight perfect bipartite matching problem $\mathcal{P}_3$, and then further relax $\mathcal{P}_3$ into the equivalent linear programming problem $\widehat{\mathcal{P}}_3$. These two transformations preserve the optimality of problem $\widehat{\mathcal{P}}$. Hence, the optimal RB allocation policy can be obtained by solving $\widehat{\mathcal{P}}_3$. The corresponding optimal client scheduling policy is then given by $\alpha_{n,t}^* = \sum_{k=1}^{K} \theta_{n,t}^{*,(k)}$, and the optimal pruning policy for each client follows from (19).

Based on the above analysis, problem $\widehat{\mathcal{P}}$ can be solved efficiently to obtain the optimal client scheduling, neural network pruning, and RB allocation policies. For clarity, the complete procedure is summarized in Algorithm 1. Specifically, Algorithm 1 requires computing the optimal pruning policies of all clients according to (19), which involves solving the pruning policy for each client and RB pair.

V. Experimental Results

In this section, we evaluate the performance of our proposed mFedDNP and the dynamic client scheduling, neural network pruning, RB allocation algorithm. If not specified, the default system settings are given as Table I.

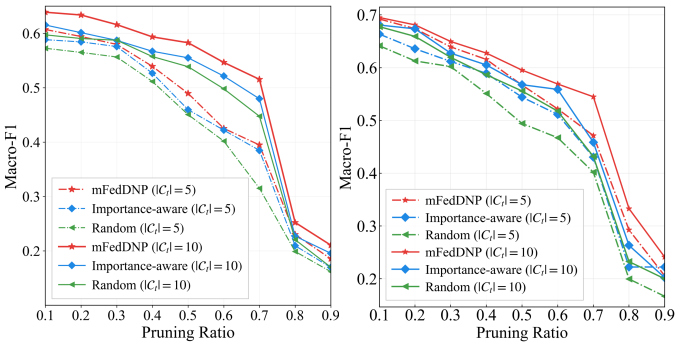
TABLE I
System Parameters

Parameter	Value	Parameter	Value
N	100	K	10
B	1MHz	ν_0	-174dBm
q_n $(\forall n \in \mathcal{N})$	4	h_0	-30dBm
ϵ	32bits	η	0.005
I	5	L_b	64

A. Implementation Details

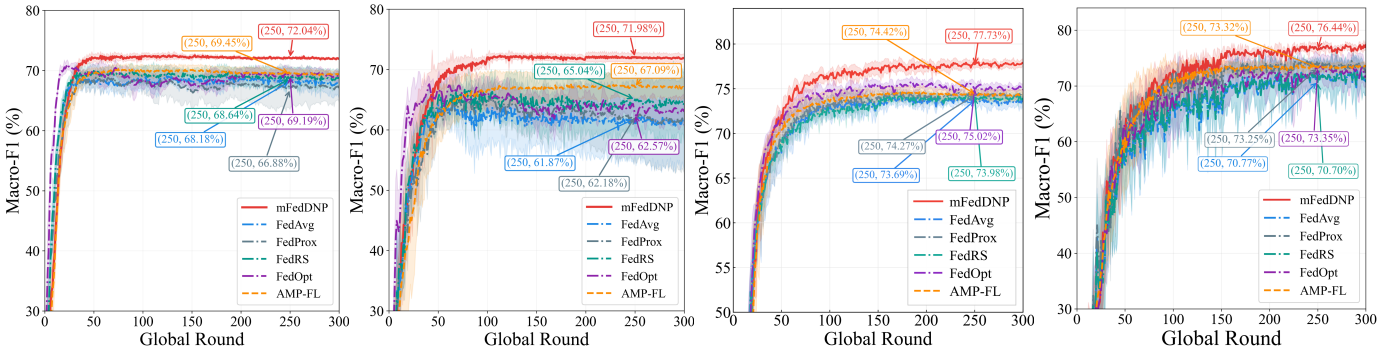
Wireless Network Setting: The simulation considers a single server with $N_r = 4$ receive antennas located at the center of a circular area with radius 500m, serving $N = 100$ randomly distributed clients with single antenna. For each client, the transmit power is randomly selected from $\{0.01, 0.02, 0.03, 0.05\}$W, and the CPU frequency is randomly selected from $\{0.5, 0.8, 1.0, 1.2, 1.6\}$GHz, and each CPU cycle can process 4 FLOPs. Following [9], [11], [42], [43], the channel gain is modeled as $h_{n,t} = h_0 \rho_n(t) d_n^{-2}$, where h_0 is the path loss constant, $\rho_n(t) \sim \exp(1)$ is the small-scale fading power gain, and d_n is the distance between the server and client n. Similar to [7], [8], the inter-cell interference δ_k at each RB k is randomly selected from the range of $[10^2 B\nu_0, 10^5 B\nu_0]$.

Multimodal Dataset and Neural Networks: We evaluate the proposed framework on two multimodal classification benchmarks, i.e., CREMA-D and UCI-HAR. CREMA-D is an audio-visual emotion recognition dataset, in which each



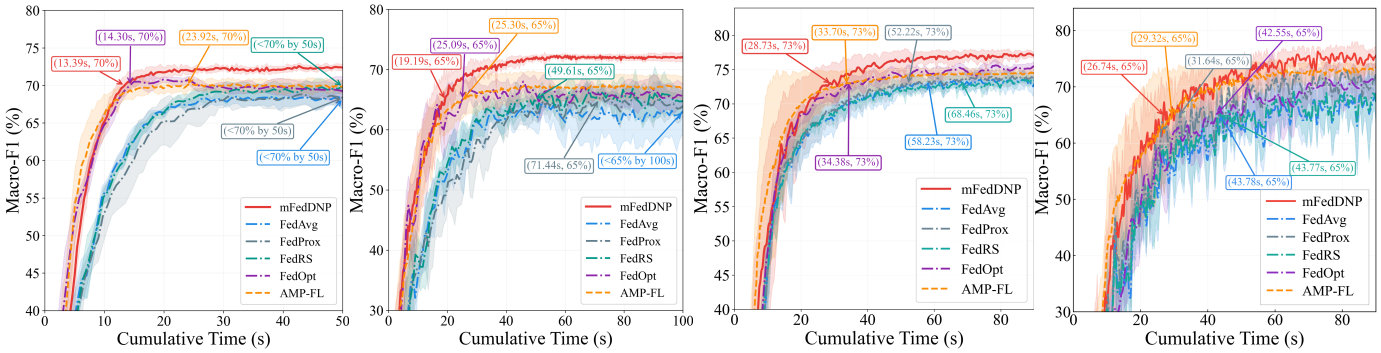
(a) CREMA-D (b) UCI-HAR

Fig. 2. Comparison of Macro-F1 achieved by different pruning methods after 200 training rounds on the CREMA-D and UCI-HAR datasets.



(a) CREMA-D, $\varphi = 5$ (b) CREMA-D, $\varphi = 0.5$ (c) UCI-HAR, $\varphi = 5$ (d) UCI-HAR, $\varphi = 0.5$

Fig. 3. Comparison of the performance of mFedDNP and benchmark methods on the CREMA-D and UCI-HAR datasets.



(a) CREMA-D, $\varphi = 5$ (b) CREMA-D, $\varphi = 0.5$ (c) UCI-HAR, $\varphi = 5$ (d) UCI-HAR, $\varphi = 0.5$

Fig. 4. Convergence speed comparison of mFedDNP and benchmark methods on the CREMA-D and UCI-HAR datasets.

TABLE II
PERFORMANCE COMPARISON BETWEEN THE PROPOSED MFEDDNP AND BENCHMARK METHODS

Dataset	Target		FedAvg		FedProx		FedRS		FedOpt		AMP-FL		**mFedDNP**		Improvement	
	$\varphi=5$	$\varphi=0.5$	$\varphi=5$	$\varphi=0.5$	$\varphi=5$	$\varphi=0.5$	$\varphi=5$	$\varphi=0.5$	$\varphi=5$	$\varphi=0.5$	$\varphi=5$	$\varphi=0.5$	$\varphi=5$	$\varphi=0.5$	$\varphi=5$	$\varphi=0.5$
CREMA-D	250	250	68.18%	61.87%	66.88%	62.18%	68.64%	65.04%	69.19%	62.57%	**69.45%**	**67.09%**	**72.04%**	**71.98%**	*2.59%*	*4.89%*
UCI-HAR	250	250	73.69%	70.77%	74.27%	73.25%	73.98%	70.70%	**75.02%**	**73.35%**	74.42%	73.32%	**77.73%**	**76.44%**	*2.71%*	*3.09%*
CREMA-D	70%	65%	>50s	>100s	>50s	71.44s	>50s	49.61s	**14.30s**	**25.09s**	23.92s	25.30s	**13.39s**	**19.19s**	*6.36%*	*23.52%*
UCI-HAR	73%	65%	58.23s	43.78s	52.22s	31.64s	68.46s	43.77s	34.38s	42.55s	**33.70s**	**29.32s**	**28.73s**	**26.74s**	*14.75%*	*8.80%*

sample contains paired audio and visual representations and the learning task is to identify the underlying emotion category. UCI-HAR is a multimodal human activity recognition dataset, where accelerometer and gyroscope signals are regarded as two complementary modalities for activity classification. For both datasets, we adopt a lightweight multimodal neural architecture composed of modality-specific multilayer perceptron (MLP) branches and a fusion network. Specifically, each modality is first processed by an individual MLP to extract a compact latent representation, and the resulting modality embeddings are then concatenated with modality-presence indicators and fed into a fusion head for final prediction. This learned-fusion architecture evaluates the staleness-aware policies motivated by the analytical model in Sections II and III. The latency expressions in Section II-B characterize the communication and computation costs of the pruned unimodal branches under the stated fusion-overhead approximation. The data distribution follows Dirichlet distribution Dir(φ). Since the Dirichlet-based data partition may induce label imbalance across clients and classes, we use Macro-F1 rather than test accuracy as the primary evaluation metric. Macro-F1 assigns equal importance to each class and therefore better reflects the balanced classification performance of different methods under heterogeneous federated settings, whereas test accuracy can be dominated by majority classes and may overestimate performance.

Baselines Setting: We select five representative baselines for comparison: 1) FedAvg [5]: the standard federated averaging baseline; 2) FedProx [44]: introduces a proximal term to improve training under heterogeneous data distributions; 3) FedRS [45]: used as a representative method for handling label imbalance and statistical heterogeneity; 4) FedOpt [46]: applies adaptive server-side optimization during global aggregation. 5) AMP-FL [17]: combines wireless-aware scheduling with adaptive model pruning. To further evaluate the effectiveness of each component in mFedDNP, we also consider four ablation variants, i.e., mFedDNP w/o MC, mFedDNP w/o DNP, mFedDNP w/o RB Allocation, and mFedDNP w/o Beamforming, together with a basic MMFL baseline. We additionally compare with Harmony [23] and MFedMC [29] to evaluate dedicated approaches to modality-heterogeneous MFL.

TABLE III
MACRO-F1 COMPARISON ON CREMA-D.

Method	$\varphi = 5$	$\varphi = 0.5$
Harmony	66.27%	60.04%
MFedMC	71.48%	59.95%
mFedDNP	**72.04%**	**71.98%**

TABLE IV
ABLATION ANALYSIS OF MFEDDNP

Method	Target		CREMA-D		UCI-HAR	
	$\varphi = 5$	$\varphi = 0.5$	$\varphi = 5$	$\varphi = 0.5$	$\varphi = 5$	$\varphi = 0.5$
Base MMFL	*68%*	*65%*	$> 60s$	$> 120s$	$> 60s$	$> 80s$
mFedDNP w/o Beamforming	*68%*	*65%*	47.86s	68.12s	$> 60s$	$> 80s$
mFedDNP w/o RB Allocation	*68%*	*65%*	16.32s	23.75s	36.80s	44.08s
mFedDNP w/o DNP	*68%*	*65%*	**14.28s**	**21.72s**	28.55s	**30.60s**
mFedDNP w/o MC	*68%*	*65%*	17.88s	28.33s	**26.14s**	32.98s
Full mFedDNP	*68%*	*65%*	**11.59s**	**19.19s**	**21.34s**	**26.70s**
Speed Up	*68%*	*65%*	***18.84%***	***11.65%***	***18.36%***	***12.75%***

B. Comparison of Pruning Approaches

Fig. 2 compares the proposed pruning strategy in mFedDNP with two benchmark pruning schemes, namely importance-aware pruning and random pruning, on the CREMA-D and UCI-HAR datasets after 200 global training rounds. It can be observed that the proposed method consistently achieves the highest Macro-F1 over a wide range of pruning ratios under both client participation settings, i.e., $|\mathcal{C}_t|= 5$ and $|\mathcal{C}_t|= 10$. This result demonstrates that the proposed staleness-aware pruning strategy is more effective in preserving task-relevant neural network regions during wireless multimodal federated learning.

As the pruning ratio increases, the Macro-F1 of all methods gradually decreases, and the performance degradation becomes particularly pronounced in the high-pruning regime. This is expected, since excessive pruning removes too many filters and neurons, thereby weakening the representation capability of the unimodal branches and the final multimodal prediction performance. Nevertheless, compared with importance-aware pruning and random pruning, the proposed mFedDNP exhibits a noticeably slower degradation trend on both datasets, indicating stronger robustness against aggressive model compression.

In addition, increasing the number of participating clients improves the performance of all pruning schemes, since more scheduled clients provide richer multimodal information for global model updating. Even in this case, however, the proposed pruning approach still remains superior to the benchmark methods, which further verifies its effectiveness. Overall, Fig. 2 confirms that the proposed pruning design achieves a more favorable accuracy-compression tradeoff and is therefore adopted in the subsequent experiments.

C. Superiority of Proposed mFedDNP

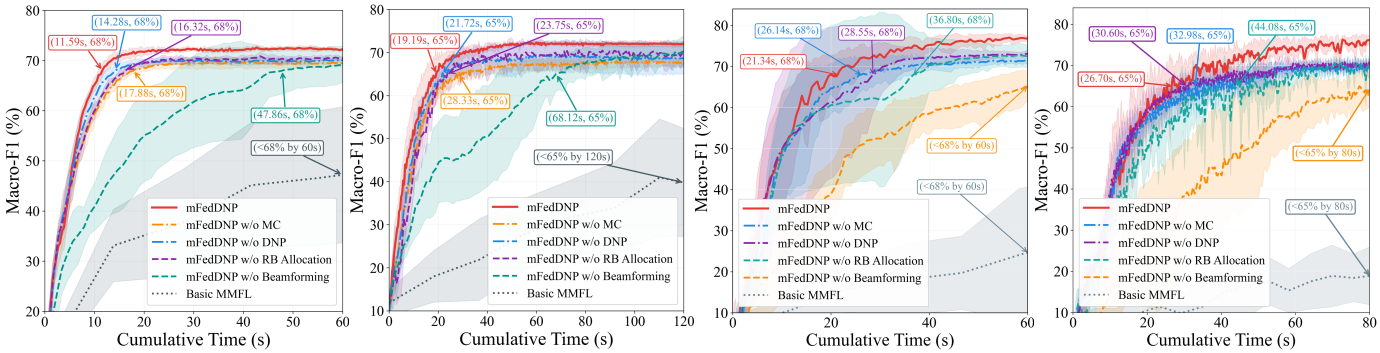
Figs. 3 and 4 together with Table II compare the proposed mFedDNP with FedAvg, FedProx, FedRS, FedOpt, and AMP-FL on the CREMA-D and UCI-HAR datasets under two Dirichlet settings. As shown in Fig. 3, mFedDNP consistently achieves the best learning performance over the whole training process on both datasets. In particular, after 250 global rounds, mFedDNP attains Macro-F1 values of 72.04% and 71.98% on CREMA-D, and 77.73% and 76.44% on UCI-HAR, under the two considered data heterogeneity settings, respectively. Compared with the strongest benchmark in each case, the proposed scheme improves the final Macro-F1 by 2.59% and 4.89% on CREMA-D, and by 2.71% and 3.09% on UCI-HAR, as summarized in Table II.

The above performance gain mainly comes from the joint design of multimodal learning and wireless-aware adaptation in mFedDNP. The proposed framework combines learned multimodal fusion with modal compensation to address modality heterogeneity and incomplete-modal updates. Moreover, compared with AMP-FL, which mainly focuses on pruning and scheduling for wireless FL, mFedDNP further exploits staleness-aware neural network pruning and storage-efficient multimodal gradient maintenance. As a result, the proposed framework preserves more informative neural network regions and produces more reliable global updates, thereby achieving a higher final Macro-F1, especially under the more challenging heterogeneous setting.

Fig. 4 further demonstrates the convergence advantage of mFedDNP in terms of cumulative training time. It can be observed that mFedDNP not only converges to a better solution, but also reaches the target performance significantly faster than all benchmark methods. Specifically, for the target levels shown in Table II, mFedDNP only requires 13.39 s and 19.19 s on CREMA-D, and 28.73 s and 26.74 s on UCI-HAR, respectively. In contrast, several conventional FL baselines require substantially longer time, and some of them even fail to achieve the target within the considered time horizon. Relative to the best competing scheme, mFedDNP reduces the required convergence time by 6.36% and 23.52% on CREMA-D, and by 14.75% and 8.80% on UCI-HAR.

Overall, the results in Figs. 3 and 4 and Table II verify the clear superiority of the proposed mFedDNP in both learning accuracy and convergence efficiency. By jointly optimizing beamforming-aided uplink transmission, client scheduling, RB allocation, and dynamic neural network pruning, mFedDNP effectively mitigates the communication bottleneck and straggler effect in wireless multimodal federated learning. Meanwhile, the modal compensation mechanism improves robustness against modality heterogeneity and partial participation. Therefore, the proposed mFedDNP provides a more favorable accuracy-efficiency tradeoff than all considered benchmark schemes.

Table III compares mFedDNP with Harmony and MFedMC on CREMA-D after 250 global rounds under $\varphi = 5$ and $\varphi = 0.5$. mFedDNP achieves 72.04% and 71.98%, respectively, exceeding Harmony by 5.77 and 11.94 percentage points and MFedMC by 0.56 and 12.03 percentage points. These results complement the benchmark and ablation studies by demonstrating higher Macro-F1 than both dedicated MFL methods in the evaluated settings.



(a) CREMA-D, $\varphi = 5$ (b) CREMA-D, $\varphi = 0.5$ (c) UCI-HAR, $\varphi = 5$ (d) UCI-HAR, $\varphi = 0.5$

Fig. 5. Performance comparison of full mFedDNP and its ablation variants on the CREMA-D and UCI-HAR datasets.

D. Ablation Studies

Fig. 5 and Table IV evaluate the contribution of each key component in mFedDNP, including modal compensation (MC), dynamic neural network pruning (DNP), RB allocation, and beamforming. It can be observed that the full mFedDNP consistently achieves the best convergence behavior and the strongest final performance on both CREMA-D and UCI-HAR under both Dirichlet settings. In contrast, the basic MMFL baseline is always the slowest and fails to reach the target within the considered time horizon. Among all ablation variants, removing beamforming causes the most severe degradation on the wireless side, since the convergence time increases to 47.86 s and 68.12 s on CREMA-D and the target cannot be reached within the considered horizon on UCI-HAR. Removing RB allocation also leads to clear performance loss, with the required time increasing from 11.59 s to 16.32 s and from 19.19 s to 23.75 s on CREMA-D, and from 21.34 s to 36.80 s and from 26.70 s to 44.08 s on UCI-HAR, which confirms the importance of wireless resource orchestration.

The comparison with mFedDNP w/o DNP and mFedDNP w/o MC further verifies the necessity of the learning-side design. Without DNP, each client has to train and upload a larger local model, which increases both computation and communication cost and slows down convergence in all cases. Without MC, the server cannot effectively compensate for missing or stale unimodal updates, and the degradation becomes more visible under the more heterogeneous setting. Compared with the strongest ablation baseline, the full mFedDNP further reduces the convergence time by 18.84%, 11.65%, 18.36%, and 12.75%, respectively. The ablations support a resource-dependent design principle: beamforming and RB allocation address low-rate uplinks, while DNP reduces computation and upload workloads when limited CPU capacity or large model size makes training costly. When communication is adequate but incomplete or outdated modality information limits learning progress, MC and staleness-aware region selection address the information bottleneck. Their joint use coordinates update delivery, model workload, and information freshness.

VI. Conclusion

In this paper, we proposed mFedDNP, a beamforming-aided wireless multimodal federated learning framework with dynamic neural network pruning. By jointly designing modal compensation, neural network pruning, client scheduling, and RB allocation, the proposed method improves both learning performance and training efficiency in heterogeneous wireless networks. We also derived a staleness-aware convergence bound and developed an efficient algorithm for the resulting optimization problem. Experimental results on CREMA-D and UCI-HAR showed that mFedDNP achieves better Macro-F1 and faster convergence than the benchmark schemes. The joint modeling of pruned-model workload and wireless transmission provides a basis for energy-aware MFL. A future extension can incorporate client-specific computation and uplink energy budgets into the staleness-based optimization, enabling battery-aware scheduling and pruning under the round-latency constraint.

Appendix

A. Proof of Theorem 1

At the edge server, the maintained gradient array $\bar{\boldsymbol{G}}_t$ satisfies

$$\begin{aligned}\hat{\boldsymbol{G}}^{(r)}_{m,t} &= \hat{\boldsymbol{G}}^{(r)}_{m,t-1} + \frac{1}{N}\sum\nolimits_{n=1}^{N} \alpha_{n,t}\beta^{(r)}_{n,m,t}\hat{\boldsymbol{G}}^{(r)}_{n,m,t} \\ &= \hat{\boldsymbol{G}}^{(r)}_{m,t-1} + \frac{1}{N}\sum\nolimits_{n=1}^{N} \alpha_{n,t}\beta^{(r)}_{n,m,t}(\tilde{\boldsymbol{g}}^{(r)}_{n,m,t} - \boldsymbol{G}^{(r)}_{n,m,t-1}) \\ &= \hat{\boldsymbol{G}}^{(r)}_{m,t-1} + \frac{1}{N}\sum\nolimits_{n=1}^{N} (\boldsymbol{G}^{(r)}_{n,m,t} - \boldsymbol{G}^{(r)}_{n,m,t-1}). \end{aligned} \tag{24}$$

At the beginning of training, the client histories $\boldsymbol{G}_{n,m,-1}$ and the server aggregates $\hat{\boldsymbol{G}}_{m,-1}$ are initialized to $\mathbf{0}$. Hence, when $t = 0$, we have

$$\begin{aligned}\hat{\boldsymbol{G}}^{(r)}_{m,0} &= \hat{\boldsymbol{G}}^{(r)}_{m,-1} + \frac{1}{N}\sum\nolimits_{n=1}^{N} (\boldsymbol{G}^{(r)}_{n,m,0} - \boldsymbol{G}^{(r)}_{n,m,-1}) \\ &= \frac{1}{N}\sum\nolimits_{n=1}^{N} \boldsymbol{G}^{(r)}_{n,m,0}. \end{aligned} \tag{25}$$

When $t = 1$,

$$\begin{aligned}\hat{\boldsymbol{G}}^{(r)}_{m,1} &= \hat{\boldsymbol{G}}^{(r)}_{m,0} + \frac{1}{N}\sum\nolimits_{n=1}^{N} (\boldsymbol{G}^{(r)}_{n,m,1} - \boldsymbol{G}^{(r)}_{n,m,0}) \\ &= \frac{1}{N}\sum\nolimits_{n=1}^{N} \boldsymbol{G}^{(r)}_{n,m,1}. \end{aligned} \tag{26}$$

Similarly, for $t > 1$, $\hat{\boldsymbol{G}}^{(r)}_{m,t} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{G}^{(r)}_{n,m,t}$. Thus, the updated global unimodal submodel is $\boldsymbol{v}^{(r)}_{m,t+1} = \boldsymbol{v}^{(r)}_{m,t} - \eta\hat{\boldsymbol{G}}^{(r)}_{m,t} = \boldsymbol{v}^{(r)}_{m,t} - \eta\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{G}^{(r)}_{n,m,t}$, we complete the proof.

B. Proof of Lemma 1

For $t \geq t'$, let $\zeta = I\eta$, we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{v}_{m,t}^{(r)} - \boldsymbol{v}_{m,t'}^{(r)}\|^2 &= \zeta^2 \mathbb{E}\|\sum_{j=t'}^{t-1} \frac{1}{NI} \sum_{n=1}^{N}\sum_{i=1}^{I} \nabla \hat{L}_n(\boldsymbol{u}_1)\|^2 \\
&\overset{(a)}{\leq} \zeta^2 (t-t') \sum_{j=t'}^{t-1} \mathbb{E}\|\frac{1}{NI} \sum_{n=1}^{N}\sum_{i=1}^{I} \nabla \hat{L}_n(\boldsymbol{u}_1)\|^2 \\
&\overset{(b)}{\leq} 3\zeta^2 (t-t') \sum_{j=t'}^{t-1} \frac{1}{NI} \sum_{n=1}^{N}\sum_{i=1}^{I} \mathbb{E}\|\nabla \hat{L}_n(\boldsymbol{u}_1) - \nabla L_n(\boldsymbol{u}_1)\|^2 \\
&+ 3\zeta^2 (t-t') \sum_{j=t'}^{t-1} \frac{1}{NI} \sum_{n=1}^{N}\sum_{i=1}^{I} \mathbb{E}\|\nabla L_n(\boldsymbol{u}_1) - \nabla L_n(\boldsymbol{u}_2)\|^2 \\
&+ 3\zeta^2 (t-t') \sum_{j=t'}^{t-1} \mathbb{E}\|\nabla L(\boldsymbol{u}_2)\|^2 \\
&\overset{(c)}{\leq} 3\zeta^2 (t-t') \sum_{j=t'}^{t-1} \frac{1}{NI} \sum_{n=1}^{N}\sum_{i=1}^{I} L^2 \underbrace{\mathbb{E}\|\boldsymbol{u}_1 - \boldsymbol{u}_2\|^2}_{A} \\
&+ 3\zeta^2 (t-t')^2 (\Phi^2 + \sigma^2) \qquad (27)
\end{aligned}$$

where $\boldsymbol{u}_1 = \boldsymbol{v}^{(r)}_{m,t-\tau^{(r)}_{n,m,t},i}$ and $\boldsymbol{u}_2 = \boldsymbol{v}^{(r)}_{m,t-\tau^{(r)}_{n,m,t}}$. Step (a) follows from Jensen's inequality, step (b) is obtained by applying the triangle inequality together with Jensen's inequality, and step (c) uses Assumptions 1, 3, and 4. We next derive an upper bound for A:

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{v}_{n,m,t,i}^{(r)} - \boldsymbol{v}_{m,t}^{(r)}\|^2 \\
&= \mathbb{E}\|\boldsymbol{v}_{n,m,t,i-1}^{(r)} - \eta \nabla \hat{L}_n(\boldsymbol{v}_{m,t,i-1}^{(r)}) - \boldsymbol{v}_{m,t}^{(r)}\|^2 \\
&\overset{(a)}{\leq} \mathbb{E}\|\boldsymbol{v}_{n,m,t,i-1}^{(r)} - \boldsymbol{v}_{m,t}^{(r)} - \eta \nabla L_n(\boldsymbol{v}_{m,t,i-1}^{(r)})\|^2 + \eta^2\sigma^2 \\
&\overset{(b)}{\leq} (\frac{1}{I-1} + 1)\mathbb{E}\|\boldsymbol{v}_{n,m,t,i-1}^{(r)} - \boldsymbol{v}_{m,t}^{(r)}\|^2 \\
&+ I\eta^2 \mathbb{E}\|\nabla L_n(\boldsymbol{v}_{m,t,i-1}^{(r)})\|^2 + \eta^2\sigma^2 \\
&\overset{(c)}{\leq} (\frac{1}{I-1} + 1)\mathbb{E}\|\boldsymbol{v}_{n,m,t,i-1}^{(r)} - \boldsymbol{v}_{m,t}^{(r)}\|^2 + 2I\eta^2 \mathbb{E}\|\nabla L_n(\boldsymbol{v}_{m,t}^{(r)})\|^2 \\
&+ 2I\eta^2 \mathbb{E}\|\nabla L_n(\boldsymbol{v}_{m,t,i-1}^{(r)}) - \nabla L_n(\boldsymbol{v}_{m,t}^{(r)})\|^2 + \eta^2\sigma^2 \\
&\overset{(d)}{\leq} (2IL^2\eta^2 + \frac{1}{I-1} + 1)\mathbb{E}\|\boldsymbol{v}_{n,m,t,i-1}^{(r)} - \boldsymbol{v}_{m,t}^{(r)}\|^2 \\
&+ 2I\eta^2 \mathbb{E}\|\nabla L_n(\boldsymbol{v}_{m,t}^{(r)})\|^2 + \eta^2\sigma^2 \\
&\overset{(e)}{\leq} (\frac{3}{2I-2} + 1)\mathbb{E}\|\boldsymbol{v}_{n,m,t,i-1}^{(r)} - \boldsymbol{v}_{m,t}^{(r)}\|^2 \\
&+ 2I\eta^2 \mathbb{E}\|\nabla L_n(\boldsymbol{v}_{m,t}^{(r)})\|^2 + \eta^2\sigma^2, \qquad (28)
\end{aligned}$$

where (a) follows from the triangle inequality together with Assumption 3, (b) is obtained by applying the triangle inequality, (c) uses the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, (d) follows from Assumption 1, and (e) holds under the condition $\zeta < \frac{1}{2L}$. Applying this inequality recursively yields

$$\begin{aligned}
&\mathbb{E}\|\boldsymbol{v}_{n,m,t,i}^{(r)} - \boldsymbol{v}_{m,t}^{(r)}\|^2 \\
&\leq \frac{(\frac{3}{2I-2} + 1)^{I-1} - 1}{\frac{3}{2I-2}} (2I\eta^2 \mathbb{E}\|\nabla L_n(\boldsymbol{v}_{m,t}^{(r)})\|^2 + \eta^2\sigma^2) \\
&\overset{(a)}{\leq} 4(I-1)(2I\eta^2\Phi^2 + \eta^2\sigma^2), \qquad (29)
\end{aligned}$$

where (a) is derived from $(\frac{3}{2I-2}+1)^{I-1} = (\frac{3}{2I-2}+1)^{\frac{3}{2}\frac{2I-2}{3}} \leq e^{\frac{3}{2}} < 5$. Substituting (29) back into (27), we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{v}_{m,t}^{(r)} - \boldsymbol{v}_{m,t'}^{(r)}\|^2 &\leq 3\eta^2 (t-t')^2 [(2R(I-1) + I^2)\Phi^2 \\
&+ ((I-1)R + I^2)\sigma^2)], \qquad (30)
\end{aligned}$$

we complete the proof.

C. Proof of Theorem 2

By the L-smoothness of the local loss function in Assumption 1, we have $\mathbb{E}[L(\boldsymbol{w}_{t+1}) - L(\boldsymbol{w}_t)] \leq \mathbb{E}\langle \nabla L(\boldsymbol{w}_t), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t\rangle + \frac{L}{2}\mathbb{E}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2$. Since both the inner-product term and the norm term can be decomposed over all parameter regions, the above inequality can be rewritten as

$$\begin{aligned}
&\mathbb{E}[L(\boldsymbol{w}_{t+1}) - L(\boldsymbol{w}_t)] \leq \mathbb{E}\langle \nabla L(\boldsymbol{w}_t), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t\rangle \\
&+ \frac{L}{2}\mathbb{E}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 \\
&= \underbrace{-\zeta \mathbb{E}\langle \nabla L(\boldsymbol{w}_t), \sum_{r=1}^{R}\sum_{m=1}^{M} \frac{1}{NI} \sum_{n=1}^{N}\sum_{i=0}^{I-1} \nabla \hat{L}_n(\boldsymbol{v}^{(r)}_{m,t-\tau^{(r)}_{n,m,t},i})\rangle}_{E_1} \\
&+ \underbrace{\frac{L}{2}\zeta^2 \sum_{r=1}^{R}\sum_{m=1}^{M} \mathbb{E}\|\frac{1}{NI} \sum_{n=1}^{N}\sum_{i=0}^{I-1} \nabla \hat{L}_n(\boldsymbol{v}^{(r)}_{m,t-\tau^{(r)}_{n,m,t},i})\|^2}_{E_2}. \qquad (31)
\end{aligned}$$

Now we bound the two terms in (31). For E_1,

$$\begin{aligned}
E_1 &\overset{(a)}{=} -\zeta \mathbb{E}\langle \nabla L(\boldsymbol{w}_t), \sum_{r=1}^{R}\sum_{m=1}^{M} \frac{1}{NI} \sum_{n=1}^{N}\sum_{i=0}^{I-1} \nabla L_n(\boldsymbol{u}_1)\rangle \\
&- \zeta \mathbb{E}\langle \nabla L(\boldsymbol{w}_t), \sum_{r=1}^{R}\sum_{m=1}^{M} \frac{1}{NI} \sum_{n=1}^{N}\sum_{i=0}^{I-1} (\nabla \hat{L}_n(\boldsymbol{u}_1) - \nabla L_n(\boldsymbol{u}_1))\rangle \qquad (32)
\end{aligned}$$

where $\boldsymbol{u}_1 = \boldsymbol{v}^{(r)}_{m,t-\tau^{(r)}_{n,m,t},i}$, (a) is due to add and subtract $\nabla L_n(\boldsymbol{u}_1)$. To bound the first term of E_1, i.e., $E_{1.1}$, we have

$$\begin{aligned}
E_{1.1} &= -\zeta \mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 \\
&+ \zeta \mathbb{E}\langle \nabla L(\boldsymbol{w}_t), \nabla L(\boldsymbol{w}_t) - \sum_{r=1}^{R}\sum_{m=1}^{M} \frac{1}{NI} \sum_{n=1}^{N}\sum_{i=0}^{I-1} \nabla L_n(\boldsymbol{u}_1)\rangle \\
&\overset{(a)}{\leq} -\frac{\zeta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 \\
&+ \frac{\zeta}{2} \sum_{r=1}^{R}\sum_{m=1}^{M} \mathbb{E}\|\frac{1}{NI} \sum_{n=1}^{N}\sum_{i=0}^{I-1} (\nabla L_n(\boldsymbol{v}_{m,t}^{(r)}) - \nabla L_n(\boldsymbol{u}_1))\|^2 \\
&\overset{(b)}{\leq} -\frac{\zeta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 \\
&+ \frac{\zeta}{2} \sum_{r=1}^{R}\sum_{m=1}^{M} \frac{1}{NI} \sum_{n=1}^{N}\sum_{i=0}^{I-1} \mathbb{E}\|\nabla L_n(\boldsymbol{v}_{m,t}^{(r)}) - \nabla L_n(\boldsymbol{u}_1)\|^2 \\
&\overset{(c)}{\leq} -\frac{\zeta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2 \\
&+ \zeta \sum_{r=1}^{R}\sum_{m=1}^{M} \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}\|\nabla L_n(\boldsymbol{v}_{m,t}^{(r)}) - \nabla L_n(\boldsymbol{u}_2)\|^2 \\
&+ \zeta \sum_{r=1}^{R}\sum_{m=1}^{M} \frac{1}{NI} \sum_{n=1}^{N}\sum_{i=0}^{I-1} \mathbb{E}\|\nabla L_n(\boldsymbol{u}_2) - \nabla L_n(\boldsymbol{u}_1)\|^2
\end{aligned}$$

$$\overset{(d)}{\leq} -\frac{\zeta}{2}\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2+\zeta\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N}L^2\mathbb{E}\|\boldsymbol{v}_{m,t}^{(r)}-\boldsymbol{u}_2\|^2 + \zeta\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{NI}\sum_{n=1}^{N}\sum_{i=0}^{I-1}L^2\mathbb{E}\|\boldsymbol{u}_2-\boldsymbol{u}_1\|^2 \tag{33}$$

where (a) follows the triangle inequality, (b) follows Jensen's inequality, (c) comes from adding and subtracting $\nabla L_n(\boldsymbol{u}_2)$, where $\boldsymbol{u}_2 = \boldsymbol{v}^{(r)}_{m,t-\tau^{(r)}_{n,m,t}}$, (d) is due to Assumption 1. To bound the second term of E_1, i.e., $E_{1.2}$, we have

$$\begin{aligned}
E_{1.2} &\overset{(a)}{=} -\zeta\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{NI}\sum_{n=1}^{N}\sum_{i=0}^{I-1}\mathbb{E}\langle\nabla L(\boldsymbol{v}_{m,t}^{(r)})-\nabla L_n(\boldsymbol{u}_2),\\
&\nabla\hat{L}_n(\boldsymbol{u}_1)-\nabla L_n(\boldsymbol{u}_1)\rangle\\
&\overset{(b)}{\leq}\frac{\zeta}{2}\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\|\nabla L(\boldsymbol{v}_{m,t}^{(r)})-\nabla L_n(\boldsymbol{u}_2)\|^2+\frac{1}{2}\zeta\sigma^2\\
&\overset{(c)}{\leq}\frac{\zeta}{2}L^2\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\|\boldsymbol{v}_{m,t}^{(r)}-\boldsymbol{u}_2\|^2+\frac{1}{2}\zeta\sigma^2
\end{aligned}\tag{34}$$

where (a) comes from adding and subtracting $\nabla L(\boldsymbol{u}_2)$, (b) is derived from Assumption 3 and the triangle inequality, (c) follows Assumption 1. Substituting (33) and (34) into (32), we have $E_1 = -\frac{1}{2}\zeta\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2+\zeta\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N}L^2\mathbb{E}\|\boldsymbol{v}_{m,t}^{(r)}-\boldsymbol{u}_2\|^2 +\zeta\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{NI}\sum_{n=1}^{N}\sum_{i=0}^{I-1}L^2\mathbb{E}\|\boldsymbol{u}_2-\boldsymbol{u}_1\|^2 +\frac{\zeta}{2}L^2\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\|\boldsymbol{v}_{m,t}^{(r)}-\boldsymbol{u}_2\|^2+\frac{1}{2}\zeta\sigma^2$. Now we turn to bound E_2:

$$\begin{aligned}
E_2 &= \frac{L}{2}\zeta^2\sum_{r=1}^{R}\sum_{m=1}^{M}\mathbb{E}\|\frac{1}{NI}\sum_{n=1}^{N}\sum_{i=0}^{I-1}\\
&(\nabla\hat{L}_n(\boldsymbol{u}_1)-\nabla L_n(\boldsymbol{u}_1)+\nabla L_n(\boldsymbol{u}_1))\|^2\\
&\overset{(a)}{\leq}\frac{L}{2}\zeta^2\sum_{r=1}^{R}\sum_{m=1}^{M}\mathbb{E}\|\frac{1}{NI}\sum_{n=1}^{N}\sum_{i=0}^{I-1}\nabla L_n(\boldsymbol{u}_1)\|^2+\frac{L}{2}\zeta^2\sigma^2\\
&\overset{(b)}{\leq} L\zeta^2\sum_{r=1}^{R}\sum_{m=1}^{M}\mathbb{E}\|\frac{1}{NI}\sum_{n=1}^{N}\sum_{i=0}^{I-1}\nabla L_n(\boldsymbol{u}_1)-\nabla L_n(\boldsymbol{v}_{m,t}^{(r)})\|^2\\
&+L\zeta^2\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2+\frac{L}{2}\zeta^2\sigma^2
\end{aligned}\tag{35}$$

where step (a) follows from the triangle inequality together with Assumption 3, and step (b) is obtained by adding and subtracting $\nabla L_n(\boldsymbol{v}_{m,t}^{(r)})$ and then applying the triangle inequality. We next bound the first term in E_2, denoted by $E_{2.1}$, as follows:

$$\begin{aligned}
E_{2.1} &\overset{(a)}{\leq} 2L\zeta^2\sum_{r=1}^{R}\sum_{m=1}^{M}\mathbb{E}\|\frac{1}{NI}\sum_{n=1}^{N}\sum_{i=0}^{I-1}(\nabla L_n(\boldsymbol{u}_1)-\nabla L_n(\boldsymbol{u}_2))\|^2\\
&+2L\zeta^2\sum_{r=1}^{R}\sum_{m=1}^{M}\mathbb{E}\|\frac{1}{N}\sum_{n=1}^{N}(\nabla L_n(\boldsymbol{u}_2)-\nabla L_n(\boldsymbol{v}_{m,t}^{(r)}))\|^2\\
&\overset{(b)}{\leq} 2L^3\zeta^2\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{NI}\sum_{n=1}^{N}\sum_{i=0}^{I-1}\mathbb{E}\|\boldsymbol{u}_1-\boldsymbol{u}_2\|^2\\
&+2L^3\zeta^2\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\|\boldsymbol{u}_2-\boldsymbol{v}_{m,t}^{(r)}\|^2
\end{aligned}\tag{36}$$

where (a) is due to adding and subtracting $\nabla L_n(\boldsymbol{u}_2)$ into $E_{2.1}$ and the triangle inequality, (b) comes from Jensen's inequality and Assumption 1. Hence, we have $E_2 \leq 2L^3\zeta^2\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{NI}\sum_{n=1}^{N}\sum_{i=0}^{I-1}\mathbb{E}\|\boldsymbol{u}_1-\boldsymbol{u}_2\|^2 +2L^3\zeta^2\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\|\boldsymbol{u}_2-\boldsymbol{v}_{m,t}^{(r)}\|^2 +\frac{L}{2}\zeta^2\sigma^2 +L\zeta^2\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2$. Hence, let $\zeta < \frac{1}{2L}$, then substituting E_1 and E_2 into (31), we have

$$\begin{aligned}
&\mathbb{E}[L(\boldsymbol{w}_{t+1})-L(\boldsymbol{w}_t)] \leq (-\frac{1}{2}\zeta+L\zeta^2)\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2\\
&+2L^2\zeta\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{NI}\sum_{n=1}^{N}\sum_{i=0}^{I-1}\mathbb{E}\|\boldsymbol{u}_2-\boldsymbol{u}_1\|^2\\
&+\frac{5}{2}L^2\zeta\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\|\boldsymbol{v}_{m,t}^{(r)}-\boldsymbol{u}_2\|^2+\frac{3}{4}\zeta\sigma^2.
\end{aligned}\tag{37}$$

According to Lemma 1, substituting (29) and (30) into (37), we have

$$\begin{aligned}
&\mathbb{E}[L(\boldsymbol{w}_{t+1})-L(\boldsymbol{w}_t)] \leq (-\frac{1}{2}I\eta+I^2L\eta^2)\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2\\
&+2IL^2RM\eta(4(I-1)(2I\eta^2\Phi^2+\eta^2\sigma^2))\\
&+\frac{5}{2}IL^2\eta[3\eta^2[(2R(I-1)+I^2)\Phi^2+((I-1)R+I^2)\sigma^2)]]\\
&\sum_{r=1}^{R}\sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N}(\tau_{n,m,t}^{(r)})^2+\frac{3}{4}I\eta\sigma^2\\
&=(-\frac{1}{2}I\eta+I^2L\eta^2)\mathbb{E}\|\nabla L(\boldsymbol{w}_t)\|^2+\gamma_1\\
&+\gamma_2\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{r=1}^{R}(\tau_{n,m,t}^{(r)})^2,
\end{aligned}\tag{38}$$

where $\gamma_1 = \frac{3}{4}I\eta\sigma^2+8IL^2RM\eta(I-1)(2I\eta^2\Phi^2+\eta^2\sigma^2)$, $\gamma_2 = \frac{15}{2}IL^2\eta^3[(2R(I-1)+I^2)\Phi^2+((I-1)R+I^2)\sigma^2)]$. According to (13), we have

$$(\tau_{n,m,t}^{(r)})^2 = (1-\alpha_{n,t}\beta_{n,m,t}^{(r)})^2(\tau_{n,m,t-1}^{(r)}+1)^2. \tag{39}$$

Note that $\alpha_{n,t}\beta_{n,m,t}^{(r)} \in \{0,1\}$, thus $(1-\alpha_{n,t}\beta_{n,m,t}^{(r)})^2 = 1-\alpha_{n,t}\beta_{n,m,t}^{(r)}$. Hence, by substituting (39) back into (38), we complete the proof.

D. Proof of Corollary 1

By Assumption 2, the gradient norm satisfies $\|\nabla L(\boldsymbol{w}_t)\|^2 \geq 2\rho(L(\boldsymbol{w}_t)-L(\boldsymbol{w}^*))$. Substituting this result into (15) and subtracting $L(\boldsymbol{w}^*)$ from both sides yields

$$\begin{aligned}
&\mathbb{E}[L(\boldsymbol{w}_{t+1})-L(\boldsymbol{w}^*)]\\
&\leq(1-I\eta\rho+2I^2L\eta^2\rho)\mathbb{E}[L(\boldsymbol{w}_t)-L(\boldsymbol{w}^*)]+\gamma_1\\
&+\frac{\gamma_2}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{r=1}^{R}(1-\alpha_{n,t}\beta_{n,m,t}^{(r)})(\tau_{n,m,t-1}^{(r)}+1)^2.
\end{aligned}\tag{40}$$

Applying the above recursion over t completes the proof.

REFERENCES

[1] C. Xie, Z. Chen, W. Yi, H. Shin, and A. Nallanathan, "Beamforming-aided wireless multimodal federated learning: Dynamic pruning and joint optimization," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*. IEEE, 2026, pp. 1770–1775.

[2] L. U. Khan, W. Saad, Z. Han, E. Hossain, and C. S. Hong, "Federated learning for internet of things: Recent advances, taxonomy, and open challenges," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1759–1799, 2021.

[3] T. Wu, K. Zhi, J. Yao, X. Lai, J. Zheng, H. Niu, M. Elkashlan, K.-K. Wong, C.-B. Chae, Z. Ding *et al.*, "Fluid antenna systems enabling 6g: Principles, applications, and research directions," *IEEE Wireless Commun.*, 2025.

[4] Z. Chen, B. Zhu, J. Wang, H. Shin, A. Nallanathan, and D. Niyato, "Network edge inference for large language models: Principles, techniques, and opportunities," *ACM Comput. Surv.*, vol. 58, no. 12, pp. 1–35, 2026.

[5] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. AISTATS*, 20–22, Apr. 2017.

[6] C. Xie, Z. Chen, W. Yi, H. Shin, and A. Nallanathan, "Distribution deviation-aware split federated learning in resource-limited wireless networks," *IEEE Trans. Commun.*, 2026.

[7] M. Chen, H. V. Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2457–2471, 2020.

[8] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, 2021.

[9] Z. Yang, M. Chen, W. Saad, C. S. Hong, and M. Shikh-Bahaei, "Energy efficient federated learning over wireless communication networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1935–1949, 2021.

[10] P. S. Bouzinis, P. D. Diamantoulakis, and G. K. Karagiannidis, "Wireless quantized federated learning: A joint computation and communication design," *IEEE Trans. Commun.*, vol. 71, no. 5, pp. 2756–2770, 2023.

[11] Z. Chen, W. Yi, Y. Liu, and A. Nallanathan, "Robust federated learning for unreliable and resource-limited wireless networks," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 9793–9809, 2024.

[12] K. Wang, Y. Ma, M. Boloursaz Mashhadi, C. H. Foh, R. Tafazolli, and Z. Ding, "Convergence acceleration in wireless federated learning: A Stackelberg game approach," *IEEE Trans. Veh. Technol.*, vol. 74, no. 1, pp. 714–729, Jan. 2025.

[13] K. Wang, Z. Ding, D. K. C. So, and Z. Ding, "Exploring age-of-information weighting in federated learning under data heterogeneity," *IEEE Trans. Wireless Commun.*, vol. 24, no. 6, pp. 5023–5036, Jun. 2025.

[14] K. Wang, Z. Ding, D. K. C. So, and Z. Ding, "Age-of-information minimization in federated learning based networks with Non-IID dataset," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8939–8953, Aug. 2024.

[15] G. Chen, J. Li, Y. Cui, Q. Wu, Y. Ni, M. Hua, and S. Lu, "IRS aided federated learning: Multiple access and fundamental tradeoff," *IEEE Trans. Mobile Comput.*, vol. 25, no. 7, pp. 10 402–10 419, Jul. 2026.

[16] Y. Jiang, S. Wang, V. Valls, B. J. Ko, W.-H. Lee, K. K. Leung, and L. Tassiulas, "Model pruning enables efficient federated learning on edge devices," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10 374–10 386, 2022.

[17] Z. Chen, W. Yi, H. Shin, and A. Nallanathan, "Adaptive model pruning for communication and computation efficient wireless federated learning," *IEEE Trans. Wireless Commun.*, vol. 23, no. 7, pp. 7582–7598, 2023.

[18] M. Kim, A. L. Swindlehurst, and D. Park, "Beamforming vector design and device selection in over-the-air federated learning," *IEEE Trans. Wireless Commun.*, vol. 22, no. 11, pp. 7464–7477, 2023.

[19] F. M. Kalarde, M. Dong, B. Liang, Y. A. E. Ahmed, and H. T. Cheng, "Power-efficient over-the-air aggregation with receive beamforming for federated learning," *IEEE Trans. Wireless Commun.*, 2025.

[20] B. Xiong, X. Yang, F. Qi, and C. Xu, "A unified framework for multimodal federated learning," *Neurocomputing*, vol. 480, pp. 110–118, 2022.

[21] J. Chen and A. Zhang, "Fedmsplit: Correlation-adaptive federated multitask learning across multimodal split networks," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2022, pp. 87–96.

[22] P. Qi, D. Chiaro, and F. Piccialli, "Fl-fd: Federated learning-based fall detection with multimodal data fusion," *Information fusion*, vol. 99, p. 101890, 2023.

[23] X. Ouyang, Z. Xie, H. Fu, S. Cheng, L. Pan, N. Ling, G. Xing, J. Zhou, and J. Huang, "Harmony: Heterogeneous multi-modal federated learning through disentangled model training," in *Proc. 21st Annu. Int. Conf. Mobile Syst., Appl. Serv.*, 2023, pp. 530–543.

[24] Q. Yu, Y. Liu, Y. Wang, K. Xu, and J. Liu, "Multimodal federated learning via contrastive representation ensemble," *arXiv preprint arXiv:2302.08888*, 2023.

[25] T. Feng, D. Bose, T. Zhang, R. Hebbar, A. Ramakrishna, R. Gupta, M. Zhang, S. Avestimehr, and S. Narayanan, "Fedmultimodal: A benchmark for multimodal federated learning," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2023, pp. 4035–4045.

[26] X. Yang, B. Xiong, Y. Huang, and C. Xu, "Cross-modal federated human activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5345–5361, 2024.

[27] R. Zhang, X. Chi, W. Zhang, G. Liu, D. Wang, and F. Wang, "Unimodal training-multimodal prediction: Cross-modal federated learning with hierarchical aggregation," *IEEE Trans. Mobile Comput.*, 2025.

[28] 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.* IEEE, 2022, pp. 1469–1478.

[29] L. Yuan, D.-J. Han, S. Wang, D. Upadhyay, and C. G. Brinton, "Communication-efficient multimodal federated learning: Joint modality and client selection," *IEEE Trans. Mobile Comput.*, 2026.

[30] H. Song, J. Wang, J. Zhou, and L. Wang, "Tackling modality-heterogeneous client drift holistically for heterogeneous multimodal federated learning," *IEEE Trans. Med. Imag.*, 2024.

[31] Y. Fan, W. Xu, H. Wang, J. Zhu, and S. Guo, "Balanced multimodal federated learning via cross-modal infiltration," *arXiv preprint arXiv:2401.00894*, 2023.

[32] Y. Fan, W. Xu, H. Wang, F. Huo, J. Chen, and S. Guo, "Overcome modal bias in multi-modal federated learning via balanced modality selection," in *Eur. Conf. Comput. Vis.* Springer, 2024, pp. 178–195.

[33] 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.

[34] H. Q. Le, M. N. Nguyen, C. M. Thwal, Y. Qiao, C. Zhang, and C. S. Hong, "Fedmekt: Distillation-based embedding knowledge transfer for multimodal federated learning," *Neural Netw.*, vol. 183, p. 107017, 2025.

[35] Y. Peng, Y. Wu, J. Bian, and J. Xu, "Hybrid federated learning for multimodal iot systems," *IEEE Internet Things J.*, 2024.

[36] W. Fan, N. Wei, A. Bazzi, R. Xi, Z. Song, Y. Li, Z. Zeng, Y. Xiu, and C. Assi, "Optimizing tracking accuracy in energy-constrained multimodal ISAC via Lyapunov-driven heterogeneous mixture-of-experts," *arXiv preprint arXiv:2604.26330*, 2026.

[37] W. Fan, N. Wei, R. Xi, A. Bazzi, Y. Xiu, C. Assi, J. Dong, and J. Jin, "Heterogeneous mixture-of-experts for energy-efficient multimodal ISAC in highly mobile networks," *arXiv preprint arXiv:2604.06697*, 2026.

[38] Y. Xiu, S. Yang, W. Lyu, P. L. Yeoh, Y. Li, and Y. Ai, "Movable antenna enabled ISAC beamforming design for low-altitude airborne vehicles," *IEEE Wireless Commun. Lett.*, vol. 14, no. 5, pp. 1311–1315, 2025.

[39] E. Diao, J. Ding, and V. Tarokh, "Heterofl: Computation and communication efficient federated learning for heterogeneous clients," *arXiv preprint arXiv:2010.01264*, 2020.

[40] A. Schrijver, *Combinatorial optimization: polyhedra and efficiency*. Springer Science & Business Media, 2003, vol. 24.

[41] M. B. Cohen, Y. T. Lee, and Z. Song, "Solving linear programs in the current matrix multiplication time," *J. ACM*, vol. 68, no. 1, pp. 1–39, 2021.

[42] C. Xie, Z. Chen, W. Yi, H. Shin, and A. Nallanathan, "Tackling class imbalance and client heterogeneity for split federated learning in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 24, no. 6, pp. 4920–4936, 2025.

[43] C. Xie, Z. Chen, W. Yi, H. Xing, C.-T. Lam, and A. Nallanathan, "Straggler-resilient hierarchical split federated learning via smashed data generation," *IEEE Internet Things J.*, pp. 1–1, 2026.

[44] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proc. Mach. Learn. Syst.*, vol. 2, pp. 429–450, 2020.

[45] X.-C. Li and D.-C. Zhan, "Fedrs: Federated learning with restricted softmax for label distribution non-iid data," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discov. Data Min.*, 2021, pp. 995–1005.

[46] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," *arXiv preprint arXiv:2003.00295*, 2020.