Adaptive Hyper-graph Aggregation for Modality-Agnostic Federated Learning

Fan Qi*, Shuai Li
Tianjin University of Technology, Tianjin, China
fanqi@email.tjut.edu.cn, lshuai@stud.tjut.edu.cn

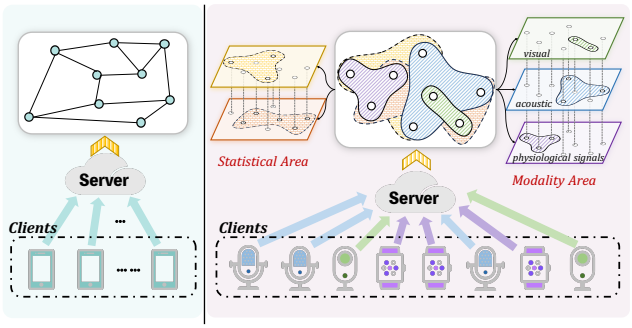
Abstract

In Federated Learning (FL), the issue of statistical data heterogeneity has been a significant challenge to the field's ongoing development. This problem is further exacerbated when clients' data vary in modalities. In response to these issues of statistical heterogeneity and modality incompatibility, we propose the Adaptive Hyper-graph Aggregation framework, a novel solution for Modality-Agnostic Federated Learning. We design a Modular Architecture for Local Model with single modality, setting the stage for efficient intra-modality sharing and inter-modality complementarity. An innovative Global Consensus Prototype Enhancer is crafted to assimilate and broadcast global consensus knowledge within the network. At the core of our approach lies the Adaptive Hyper-graph Learning Strategy, which effectively tackles the inherent challenges of modality incompatibility and statistical heterogeneity within federated learning environments, accomplishing this adaptively even without the server being aware of the clients' modalities. Our approach, tested on three multimodal benchmark datasets, demonstrated strong performance across diverse data distributions, affirming its effectiveness in multimodal federated learning.

1. Introduction

Propelled by its capacity to enable collaborative model training across decentralized data sources while upholding privacy preservation, federated learning (FL) has rapidly gained traction as a burgeoning and propitious research paradigm [1, 21, 26, 32, 39, 51]. Statistical heterogeneity caused by non-IID data across clients represents the primary challenge for federated learning algorithm research, spurring an abundance of studies centered around this challenge [9, 11, 14, 30, 38, 42].

To better handle this heterogeneity, numerous *client clustering* techniques [15, 28, 31, 48] create multiple global models by clustering clients with similar data distribution.



(a) Unimodal Collaboration Graph (b) Multimodal Collaboration Hypergraph

Figure 1. On the left, the traditional approach emphasizes parameter similarity for collaboration. On the right, our method utilizes a complex hypergraph to adaptively learn the correlation among clients, tackling modality incompatibility and statistical heterogeneity.

Typically, client similarity is evaluated by either comparing local model weights [5, 31] or loss values from different cluster models [16, 25]. However, *client clustering*-based methods lack flexibility in handling diverse data heterogeneity, as they do not specifically determine appropriate client pairs and collaboration intensity levels.

Another perspective for helping clients collaborate is to construct a dynamic graph for server aggregation [17, 27, 41]. As shown in Fig.1(a), this collaboration graph exclusively connects clients that mutually benefit from each other's data, thereby optimizing the efficiency of the learning process. In this graph, each node symbolizes an individual client's personalized model. The edges, on the other hand, represent the collaboration intensity between pairs of models, with weights that are dynamically updated in each communication round. Specifically, the server constructs a collaboration graph that emphasizes model similarity correlations, which is then utilized to generate an aggregated model for each client. On the client side, local models are optimized by balancing empirical task-driven loss and their similarity to the server's aggregated model, enhancing both individual and collaborative performance in FL.

*Corresponding author.

In a real-world FL scenario, as illustrated in Fig.1(b), visual, acoustic, and physiological signals are captured separately using different equipment. Growing interest in Multimodal Federated Learning (MFL) [2, 6, 7, 40, 44, 52] reflects its superior ability over traditional Federated Learning (FL) to exploit cross-client data modality heterogeneity, enhancing collective intelligence and broadening adaptive learning across diverse applications. However, it also spotlights challenges [12, 24, 37] such as *modality incompatibility*, where varying local data modalities result in divergent client distributions, revealing the limitations of conventional single-modality frameworks. The presence of multiple dimensions of client divergence, including ***i) modality incompatibility*** and ***ii) statistical heterogeneity***, poses challenges in identifying a unified global objective. In the absence of explicit modality indicators, the server is challenged with the dilemma of accurately clustering model parameters, a predicament that is further exacerbated when the architectures of modality-specific models remain identical across clients. **Modality-agnostic federated learning**, in particular, represents a promising but insufficiently explored avenue that could significantly enhance the capabilities of distributed learning architectures. As illustrated in Fig.1(b), a critical question emerges: **How to construct an adaptive aggregation collaboration graph on the server side that effectively considers both statistical and modality aspects?**

To tackle the above-mentioned challenges, we introduce the Adaptive Hyper-graph Aggregation for Multimodal Federated Learning (**HAMFL**) to explore relationships among heterogeneous clients, facilitating knowledge transfer and sharing to build more robust global models. **On the client side**, we introduce a modular local model structure specifically designed to accommodate the rich tapestry of multimodal data inherent to diverse client environments, structured as follows: ① The *Modality-Specific Module* is optimized to be shared with clients that collect similar types of data (*e.g.*, all visual, all auditory, etc.). ②The *Modality-Shared Module* is meant to distill and utilize the information that is generalizable across different modalities of data. ③ The *Personalized Interaction Module* adapts shared knowledge to the specific context of each client, aligning with the unique data characteristics and behavioral patterns of their specific environment.

On the server side, we introduce the Global Consensus Prototype Enhancer, designed for the key goal of efficiently assimilating and integrating knowledge from a wide range of clients. This Consensus Prototype is then distributed back to the clients to guide and calibrate their local Modality-Shared modules, aligning them with a broader spectrum of knowledge validated by the public dataset. Furthermore, we introduce a novel adaptive hypergraph aggregation strategy that intelligently identifies and models the intricate inter-client relationships within the federated network. Central to our approach is the use of local models as nodes, with hypergraph initialization executed via k-means clustering on the Modality Speculative Domain and Distributional Speculative Domain. We integrate a hypergraph that adaptively learns a collaboration matrix, with its learning process constrained by validations performed on public datasets. To ensure a balance between representation accuracy and computational efficiency, we employ a Hypergraph Diffusion Neural Network [36] in our framework.

To summarize our key contributions, we outline them as follows:

1) Our study pioneers the exploration of modality-agnostic federated learning, employing a comprehensive analytical approach to model the dynamic relationships among clients with diverse data modalities. 2) We develop an innovative Global Consensus Prototype Enhancer designed to assimilate and disseminate global consensus knowledge effectively across the network. 3) At the heart of our approach is the Adaptive Hyper-graph Learning Strategy, tailored for multimodal client adaptive aggregation on the server side. This strategy adeptly addresses issues of modality incompatibility and statistical heterogeneity. 4) Through extensive analysis across three multimodal benchmark datasets, our method has demonstrated exceptional performance in both emotion and action recognition tasks. These tasks are evaluated from first-person and third-person viewpoints under various data heterogeneity conditions, solidifying our approach's efficacy.

2. Related work

2.1. Multimodal Federated Learning

In MFL, two main configurations are discernible: Horizontal MFL, where clients house multi-modal samples, and Vertical MFL, characterized by unique or minimally overlapping modalities among clients. Our investigation pivots to Vertical MFL, emphasizing distinct, single-modality possession per client and accentuating inter-client modal disparities. Yang *et al.* [40] present the Feature-Disentangled Activity Recognition Network (FDARN) for cross-modal federated human activity recognition. Leveraging five adversarial training modules, their approach discerns both modality-agnostic attributes and modality-specific discriminatory traits of clients, outperforming prevailing personalized federated learning strategies. Zang *et al.* [49] introduce a hierarchical aggregation approach, consolidating local encoders by client-held modality types and employing attention mechanisms to synchronize decoder weights independent of data modality. Zhao *et al.* [50] segment the local network into five components, either for modality-centric aggregation for homogeneous modality clients or broad aggregation for the entire clientele. Recent advancements

[6, 40] disentangle the local model into two distinct components: modality-agnostic weights that are shared on the server side across all clients and modality-specific weights shared among subsets of clients with homogeneous data modalities. This method inherently assumes that the server can identify the modality association of each model parameter, which conflicts with the principles of privacy preservation. Such an assumption conflicts with federated learning's confidentiality principles. To mitigate the risks outlined, we introduce an adaptive hypergraph aggregation technique, allowing the server to deduce client parameter modalities via a public dataset.

2.2. Hypergraphs Neural Networks

A pioneering development within Topological Neural Networks is the use of Hypergraphs (HGs) [3, 13, 19], which transcend traditional graph structures by employing hyperedges to encapsulate multi-node connections, facilitating the representation of complex set-type relationships found in datasets spanning semantic analysis to network systems. Significant advances include Jiang *et al.*'s dynamic hypergraph neural networks [19], Yi and Park's time-series hypergraph models [43], Huang *et al.*'s message-passing interpretations in UniGNN [18], Chien *et al.*'s AllSet framework [8] integrating set functions into hypergraph learning, and Wang *et al.*'s ED-HNN [36], which innovatively approximates hypergraph diffusion processes. These advancements highlight the trajectory towards increasingly nuanced and topologically aware neural network models capable of addressing the intricate relationships inherent in complex data.

Our method innovatively applies the concept of hypergraphs to the realm of federated learning, specifically to enhance the process of model aggregation. Unlike traditional graph-based methods, which consider pairwise relations between nodes, hypergraphs allow for hyperedges that can connect multiple nodes simultaneously, capturing complex high-order interactions.

3. Methods

3.1. Preliminary

Personlized Federated Learning based on Graph. In federated settings characterized by K clients, a collaboration graph $G(\mathcal{V}, W)$ is defined to encapsulate the collaborative relationships among clients. This graph, introduced in [41], consists of a node set $\mathcal{V} = \{c_1, c_2, \ldots, c_K\}$, representing all clients, and an adjacency matrix $W \in \mathbb{R}^{K\times K}$, where each element (i, j) quantifies the extent of collaboration between the ith and jth clients. The central focus of this research is the optimization problem expressed as:

$$\min_{\{\theta_i\},W} \sum_{i=1}^{K} p_i \left(F_i \left(\sum_{j=1}^{K} W_{ij}\theta_j \right) - \lambda R(\theta_i, \theta_g) \right) \quad (1)$$

In the objective function, the first term models the empirical loss at each client after collaboration, where $\theta_i = \sum_{j=1}^{K} W_{ij}\theta_j$ is the collaborated (aggregated) model at the ith client. The second term is a regularization term quantifying the divergence between the local model parameters θ_i and the global model parameters θ_g. p_i is the weighting coefficient, which is the relative dataset size. λ is the regularization coefficient, controlling the trade-off between the global and local models.

3.2. Problem Decomposition in FL architecture

The server, equipped with a publicly available multimodal dataset D_{com}, typically sources this data from the internet, reflecting practical scenarios. We assume uniform modal distribution across each client's local data, with M representing the total number of modalities. To handle these constraints, we decompose the original problem (1) into two parts: on the server side, we learn the aggregation matrix W to aggregate local models by training a hypergraph model; on the client side, we additionally design the contrast loss function to learn modal sharing knowledge.

Solving W at the sever side. Since the server does not have access to the local dataset to evaluate the loss value of the client, we introduce a multimodal public dataset and a hypergraph network to implement the automatic aggregation strategy. The server obtains a kind of hyperedge setting $\{X, \mathcal{V}, \mathcal{E}\}$ among clients by a specific method, and the hypergraph network learns to obtain the model aggregation matrix W by a loss function. Then, the optimization for the server side is:

$$\begin{aligned} &\min_{W} \frac{1}{K} \sum_{i}^{K} (\mathcal{L}(\sum_{j}^{K} W_{ij}\theta_j, D_{com})), W = g(X, \mathcal{V}, \mathcal{E}) \\ &s.t. \sum_{j} W_{ij} = 1, \forall i; W_{ij} >= 0, \forall i, j. \end{aligned} \quad (2)$$

where $g(\cdot)$ is the server-side hypergraph model. θ_j is the model parameter of the j-th client. K is the total number of clients. Here, $\{\sum_j W_{ij}\theta_j\}_i^K$ aggregates to obtain new local models, updates the hypergraph network with the effects in the public dataset, and delivers the updated models to the client. More details will be explained in section 3.5.

Solving θ_i at the client side. Since the local client only has private data for its own modality, we design a Global Modality-Shared Prototype P_{global}, knowing that the client learns the knowledge features shared by the modality. The

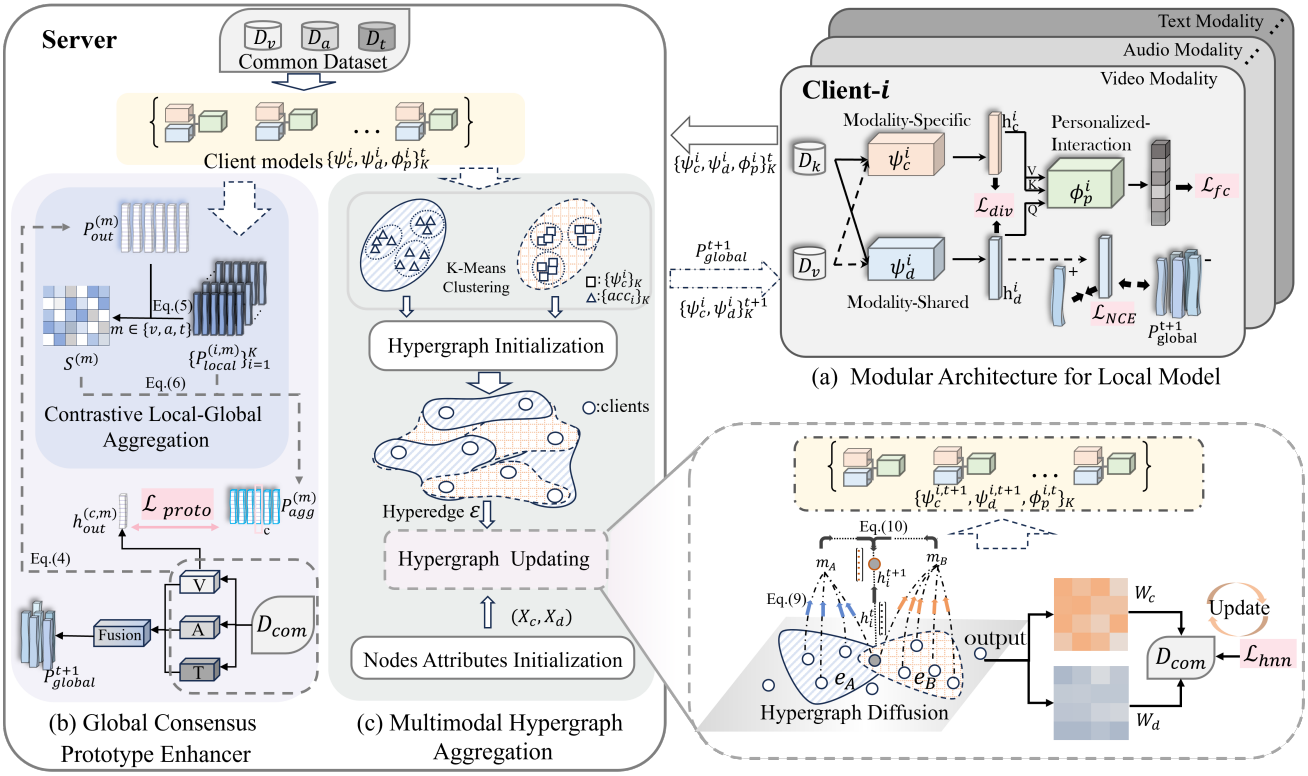


Figure 2. The network architecture of the proposed framework. Client side: (a) Modular Architecture for Local Model. Server-side: (b) Global Consensus Prototype Enhancer and (c) Multimodal Hypergraph Aggregation.

optimization for the i-th client is:

$$\min_{\theta_i} H_i(\theta_i) = \mathcal{L}(\theta_i, D_i) + \lambda \mathcal{L}'(\theta_i, P_{gloal}, D_{com}) \quad (3)$$

where the first term minimizes empirical task-driven loss to pursue local model utility, and the second term serves as a kind of prototypical regularized loss function to guide the local model in learning the modal shared feature space.

3.3. Modular Architecture for Local Model

Refer to Fig.2 for our modular approach to local models in a multimodal federated environment, comprising the **Modality-Specific module** ψ_c, **the Modality-Shared module** ψ_d, and **the Personalized Interaction module** ϕ_p. Each module serves a distinct function: ψ_c captures unique characteristics within a single modality, ψ_d extracts common features across all modalities, and ϕ_p integrates these features to enhance multimodal representation. A divergence loss, inspired by domain separation networks [4], ensures distinct feature spaces, while the Global Consensus Prototype (Section 3.4) and InfoNCE loss ensure shared feature consistency across modalities. For detailed formulations of the loss functions and their contributions to the learning process, see the supplementary materials. During training, local models are fine-tuned using a public dataset to capture cross-client generic knowledge, eschewing the direct application of server-derived prototypes to accommodate individual client characteristics. We utilize a cross-entropy loss function for the output of multimodal attention ϕ_p(detailed in the supplementary materials) to refine the integration and enhance the model's performance.

3.4. Global Consensus Prototype Enhancer

It is very critical to learn a robust and global consensus multimodal prototype across heterogeneous clients. Therefore, we introduce the Global Consensus Prototype Enhancer as shown in Fig.2(b).

Multimodal Prototype Initialization: Initially, the server trains a multimodal fusion network on a common dataset D_{com}, incorporating an attended fusion module tailored for the specific task with a cross-entropy loss function $\mathcal{L}_{CE}$. It generates the initial global model prototype for each modality as follows:

$$p_{out}^{(c,m)} = \frac{1}{|D_m^c|} \sum_{x_j \in D_m^c} f_m(x_j) \quad (4)$$

where D_m^c is the set of samples in the common dataset with

modality m and activity label c. $f_m(\cdot)$ is the m-modality module of the multimodal fusion network.

Contrastive Local-Global Aggregation: Upon receiving the Modality-Shared Modules ψ_d from clients, the server first produces the c-th class local prototype on D_{com}, labeled as $\left\{p_{local}^{(i,c,m)}, m \in \{v,a,t\}\right\}$. Subsequently, it calculates the score of each local prototype using the Info-NCE loss:

$$s^{(i,c,m)} = log\frac{exp([p_{local}^{(i,c,m)}]^\top \cdot p_{out}^{(c,m)})}{\sum_{j=1}^{K} 1_{j\neq i} exp([p_{local}^{(j,c,m)}]^\top \cdot p_{out}^{(c,m)})} \quad (5)$$

We use softmax for normalization, and the aggregated prototype for class c of modality m is formulated as follows,

$$p_{agg}^{(c,m)} = \sum_{i=1}^{K} softmax(s^{(i,c,m)}) p_{local}^{(i,c,m)} \quad (6)$$

Knowledge Transfer: In the final phase, the server model distills knowledge from the clients by minimizing the ℓ^2 loss between the global and aggregated prototypes for each modality and class:

$$\mathcal{L}_{proto}^{m} = \frac{1}{|D_m|} \sum_{x_j \in D_m} \left\| h_{out}^{(c,m)} - p_{agg}^{(c,m)} \right\|^2 \quad (7)$$

where $h_{out}^{(c,m)}$ is the intermediate output of the model: $h_{out}^{(c,m)} = f_m(x_j)$. The final loss for the whole multimodal fusion network on D_{com} is $\mathcal{L}_{CE} + \lambda \sum_{m\in v,a,t} \mathcal{L}_{proto}^{m}$. The intermediate multimodal output features of the multimodal fusion network will eventually be averaged and aggregated to obtain the global consensus prototype P_{global}^{t+1}.

3.5. Multimodal Hypergraph Aggregation

Our approach confronts the unique complexities of Heterogeneous Federated Learning (HFL), where each client's dataset is modality-specific and confidential. Unless previous multimodal federated learning approaches [40, 49, 50], this work diverges from the norm by not assuming server awareness of the clients' data modalities. To design an adaptive aggregation strategy with challenges of modality incompatibility and statistical heterogeneity, we design two domains: the Distributed Speculative Domain and the Modality Speculative Domain. Complementing this approach, we integrate Hypergraph Neural Networks (HNN) for the efficient and optimal aggregation of model parameters.

Nodes Attributes X Initialization: Each vertex in the hypergraph is characterized by the client's Modality-Specific and Modality-Shared module parameters. We expect the hypergraph network to learn the associations between clients through the parameters. The initial setting of the vertex's feature set X is:

$$X = PCA(\Theta_1, \Theta_2 ... \Theta_k) \quad (8)$$

where Θ_k is a vector of model parameters for k-th client. PCA is the Principal Component Analysis, and PCA can greatly reduce the computational effort of training without losing the effect. X_c and X_d are obtained by entering the parameters of each client Modality-Specific and Modality-Shared module into the above equations.

Hyperedge $\mathcal{E}$ Initialization: In our approach to structuring the hypergraph for Heterogeneous Federated Learning, we design hyperedges based on two key aspects:*1) Modality Speculative Domain*: We conduct a k-means clustering process applied to the Modality-Specific parameters $\{\Theta_c^i\}^K$. Through this process, we aim to identify patterns and relationships within the Modality-Specific aspects of the clients' data. Suppose the clustering results in μ_1 distinct clusters. Correspondingly, we define a set of hyperedges $\mathcal{E} = \{e_1, ..., e_{\mu_1}\}$, with each hyperedge representing a cluster. These hyperedges encapsulate the intrinsic modality-based connections within the data. *2) Distributional Speculative Domain*: This domain addresses data statistics by focusing on the accuracy across public datasets using models from all clients, based on the principle that more samples per class lead to better performance for that class. Here, a similar k-means clustering process is employed, but it is based on the accuracy metrics. Assuming this process yields μ_2 clusters, we expand our hyperedge set to include these new clusters, resulting in a combined hyperedge configuration $\mathcal{E} = \{e_1, ..., e_{\mu_1+\mu_2}\}$.

By strategically employing the Modality Speculative Domain and the Distributional Speculative Domain for hyperedge design, our hypergraph structure proficiently encapsulates the statistical characteristics of the data, while effectively addressing the challenges of modality incompatibility. This approach enhances the adaptive aggregation strategy adept at handling complex, multimodal data scenarios.

Hypergraph Updating Process: We utilize NN-parameterized Equivariant Hypergraph Diffusion Neural Operators [36], proven to excel in node label prediction on heterophilic hypergraphs where hyperedges conglomerate nodes from distinct classes. These operators adeptly facilitate dynamic interactions between unimodal and distributed domains by enabling effective diffusion of information across diverse higher-order relations. The diffusion process is as follows:

$$m_e^{(t)} = \sum_{v\in e} \varpi(h_v^{(t)}), \forall e \in \mathcal{E} \quad (9)$$

$$h_v^{(t+1)} = \varrho(h_v^{(t)}, \sum_{e:v\in e} \rho(h_v^{(t)}, m_e^{(t)})), \forall v \in \mathcal{V} \quad (10)$$

where h_v is the (latent) features of node $v \in \mathcal{V}$, and m_e is the (latent) features of hyperedge $e \in \mathcal{E}$. ϖ, ρ, ϱ are multilayer perceptions (MLPs).

The hypergraph network is utilized to derive the latest

node features ζ, which are crucial for aggregating client model parameters. The weights for aggregating the client models are given by the matrix W, with the elements defined as:

$$W_{ij} = \frac{\exp(cos(\zeta_i, \zeta_j))}{\sum_j 1_{(i,j)\in e} \exp(cos(\zeta_i, \zeta_j))} \tag{11}$$

where ζ_i denotes the latent feature of the client node i, and $(i, j) \in e$ signifies that clients i and j belong to the same hyperedge. Subsequently, the model parameters are aggregated according to $\psi_i^{t+1} = \sum_{j=1}^{K} W_{ij}\psi_j^t$.

The hypergraph network is employed independently to process the Modality-Specific ψ_c and Modality-Shared ψ_d parameters, yielding W_c and W_d, respectively. Given that the server lacks access to clients' private data, it relies on a public dataset to optimize the loss function of the hypergraph network. This setup enables the server to utilize the aggregated client models to assess accuracy on the public dataset. Now each local model is an ensemble of ψ_c^{t+1}, ψ_d^{t+1}, and ϕ_p^t, along with their respective final layers tasked with classification.

The loss function employed to update the hypergraph network is given by:

$$\mathcal{L}_{\text{hnn}} = \frac{1}{K}\sum_{i=1}^{K}(1 - \delta^{\text{acc}_i - 1}) \tag{12}$$

where acc_i is the accuracy of the updated i-th client on D_{com}, and δ is a constant. Through the exponential term, marginal precision is amplified to incentive the training of a superior collaboration graph.

4. Experiment

4.1. Implementation details

Datasets. Our proposed method is rigorously evaluated using three renowned multimodal federated datasets: EPIC-Kitchens [10], UCF-101 [33], and MELD [29], each offering distinct challenges and data modalities. For EPIC-Kitchens, encompassing 97 action classes, involves extracting both video and audio features from each dataset instance, serving as model inputs. In the case of UCF-101, which includes 101 action classes, our approach entails extracting video features and optical flow data to facilitate action recognition. MELD, with its 7 emotion classes, presents a more complex scenario requiring the fusion of video, audio, and text features for accurate emotion recognition.

Data heterogeneity. Following prior work [20, 23, 35, 46, 47], we simulate the non-IID data distributions using Dirichlet distributions, where a smaller α value corresponds to more severe data heterogeneity. For evaluation, we adopt top-1 accuracy on the above three datasets to assess the performance of our method under different α.

Baselines. We consider various state-of-the-art solutions against non-IID data distribution in the context of federated learning. Specifically, we compare with the following approaches: the vanilla aggregation strategy FedAvg [26]; based on the abstract class prototypes FedProto [34]; learnable aggregation weights FedLAW [22]; by learning a collaboration graph pFedGraph [41]; employ a dynamic and multi-view graph structure FedMSplit [6]; feature-disentangled activity recognition network FDARN [40]; Contrastive Representation Ensemble and Aggregation for Multimodal CreamFL [45]. We also compare with a baseline SingleSet, which trains a local model for each client without using FL. For all baselines, we use the publicly released code.

Implementation details. For the three datasets in our experiment, there are 4 unique modalities (i.e., video, optical flow, audio, and text). To facilitate the fair comparison with existing methods, we first extract the raw features for different modalities. Then, the raw features will be input into our model or baselines to conduct the tasks. It is worth noting that using the same dimension features of different modalities is not a mandatory requirement of our method in practice.

Our model and baselines are all trained with SGD optimizer, where the weight decay is set to 1e-5, and the momentum is set to 0.9. On the Epic-Kitchens, the learning rate η of the local client is set to 0.001, and the batch size is set to 64. On the other two datasets, the learning rate η and the batch size are set to 0.01 and 32, respectively. In addition, the learning rate η^h in the HNN model is set to 0.01, and the learning rate η^f in the multimodal fusion model is set to 0.02. δ is fixed at 64. Set λ_1 and λ_2 to 0.4 and 0.2 respectively. On all three datasets, the number of local epochs is set to 1, and the number of communication rounds T is 200. For the SingleSet baseline, the number of local epochs is set to 200. Unless explicitly specified, other hyper-parameters of each baseline are tuned within the range provided by the authors, and the best results are reported. Following [40], both the Modality-Specific and Modality-Shared modules are implemented as two-layer perceptrons with the activation function of ReLU, where the dimension of the hidden layer and the output dimension d are set to 1024 and 512. Further details are provided in the Supplementary Material.

4.2. Performance Overview

Comparison with existing methods. We report the overall performance on three datasets in Tab.1. Our model slightly outperforms the state-of-the-art methods by 0.4% on V-F action recognition and shows a notable improvement of 1.2% and 1.0% in V-A action recognition and A-V-T emo-

Methods	EPIC-Kitchens			UCF-101			MELD			
	Audio	Video	Avg.	Video	Flow	Avg.	Audio	Video	Text	Avg.
SingleSet	42.1(±0.39)	28.6(±0.52)	35.4(±0.59)	61.7(±1.14)	60.4(±1.15)	61.1(±1.22)	47.3(±1.16)	48.1(±1.62)	48.9(±1.05)	48.1(±1.81)
FedAVG [26]	41.8(±0.23)	29.7(±0.57)	35.8(±0.62)	67.8(±1.11)	66.3(±0.94)	67.1(±1.27)	48.5(±1.37)	50.8(±1.22)	49.3(±1.24)	49.5(±1.32)
FedProto [34]	42.7(±0.51)	31.2(±0.56)	36.9(±0.32)	69.7(±1.38)	69.1(±1.17)	69.4(±1.73)	49.3(±1.41)	52.7(±1.23)	51.5(±1.76)	51.2(±1.71)
FedLAW [22]	42.9(±0.49)	32.5(±0.32)	37.7(±0.65)	71.2(±0.95)	68.4(±1.22)	69.8(±1.48)	50.1(±1.02)	53.2(±1.25)	52.3(±1.74)	51.9(±1.45)
pFedGraph [41]	43.6(±0.53)	32.3(±0.51)	37.9(±0.77)	70.4(±1.34)	68.3(±1.25)	69.4(±1.24)	49.2(±1.67)	53.6(±1.26)	52.5(±1.43)	51.8(±1.53)
FedMSplit [6]	45.2(±0.61)	33.7(±0.37)	39.5(±0.42)	72.8(±1.56)	70.7(±1.48)	71.7(±1.37)	51.2(±1.25)	54.5(±1.07)	52.7(±1.64)	52.8(±1.56)
FDARN [40]	47.3(±0.55)	34.1(±0.74)	40.7(±0.64)	73.5(±1.05)	71.8(±0.87)	72.7(±1.15)	51.4(±1.47)	54.9(±1.35)	53.1(±1.16)	53.1(±1.39)
CreamFL [45]	46.9(±0.52)	33.8(±0.62)	40.4(±0.59)	73.3(±1.32)	**72.6(±1.21)**	72.9(±1.62)	—	—	—	—
Ours	**48.1(±0.57)**	**35.2(±0.41)**	**41.6(±0.53)**	**74.1(±0.91)**	72.4(±1.17)	**73.3(±1.24)**	**51.8(±1.22)**	**55.7(±1.07)**	**54.8(±1.37)**	**54.1(±1.43)**

Table 1. Performance of our HAMFL and other baseline methods on three datasets. For all methods, the non-iid of the data is set to $\alpha = 1$, and ten clients are set for each modality.

Methods	EPIC-Kitchens		
	α=0.2	α=0.5	α=1
SingleSet	26.4(±0.58)	30.6(±0.63)	35.4(±0.59)
FedAVG	28.3(±0.69)	32.5(±0.75)	35.8(±0.62)
FedProto	29.2(±0.74)	33.5(±0.37)	34.9(±0.43)
FedLAW	31.2(±0.42)	32.9(±0.59)	35.4(±0.31)
pFedGraph	31.8(±0.86)	34.2(±0.62)	37.9(±0.77)
FedMSplit	33.9(±0.71)	36.5(±0.59)	39.5(±0.42)
FDARN	34.1(±0.61)	37.6(±0.38)	40.7(±0.64)
CreamFL	33.6(±0.79)	37.1(±0.73)	40.4(±0.59)
Ours	**35.2(±0.65)**	**39.4(±0.46)**	**41.6(±0.53)**

Table 2. Performance w.r.t data heterogeneity.

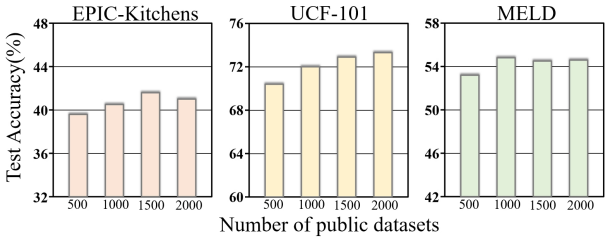


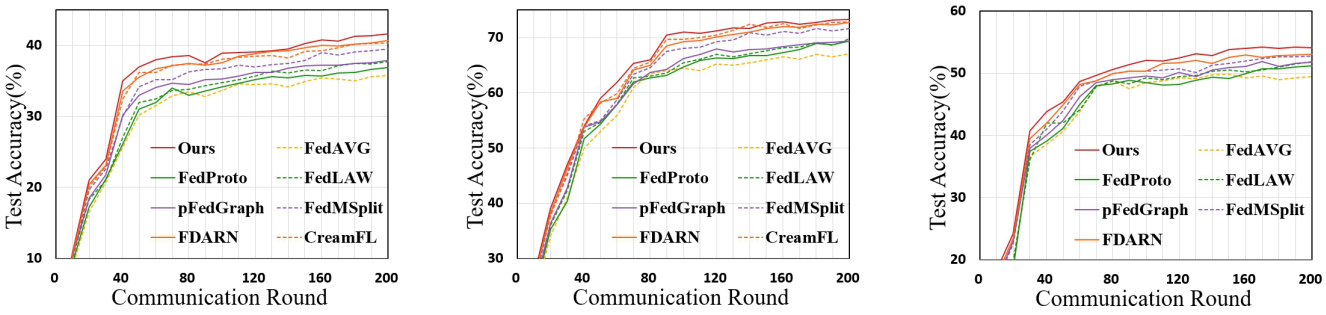
Figure 3. The impact of the number of public datasets on the performance of three datasets.

tion recognition. On UCF-101, for the flow modality, our method cannot outperform the FDARN. The sparse nature of optical flow data often leads to suboptimal aggregation results following the dimensionality reduction of network parameters. Conversely, CreamFL directly transmits the optical flow prototype, thereby avoiding the associated challenges of reprocessing sparse data. Overall, our findings indicate that: i) Our proposed method, along with other MFL approaches like FedMSplit, FDARN, and CreamFL, consistently outperforms traditional federated learning baselines across all three datasets. This underscores the notion that modality incompatibility among client modalities poses significant challenges in model learning within federated systems. ii) Notably, our method operates under modality-agnostic conditions, yet its performance closely matches or even surpasses that of modality-specific methods, demonstrating its robustness and effectiveness.

Impact of Data Heterogeneity. Demonstrating our method's robustness in federated learning, we conduct extensive experiments to assess performance under varying data heterogeneity α. Specifically, we evaluate $\alpha = 0.2$, 0.5, and 1 on three datasets, with results detailed in Tab.2. Our method significantly boosts the convergence, stabilizes training, and brings considerable performance improvement compared with previous approaches. Specifically, with heterogeneity value α =0.2, 0.5, and 1, our method has relative improvement over the FedAvg on the EPIC-Kitchens dataset by 24.4%, 21.2%, and 16.2%, respectively.

Number of Public Datasets. We investigate the impact of varying-sized public datasets on model performance and the reliability of inference outcomes. As shown in Fig.3, for the UCF-101 dataset, moderately increasing the size of the public dataset facilitates a performance improvement. A broader range of data samples allows the model to learn more comprehensive data features, enabling more accurate computation of the Global Modality-Shared Prototype and inferring the data distribution of the clients. Contrary to the above, larger public datasets adversely affect the performance of the EPIC-Kitchens and MELD. This trend is due to the highly imbalanced data distribution (i.e., long-tail distribution) inherent in the two datasets. In such cases, an overreliance on public data intensifies biases towards majority categories, consequently detrimentally impacting overall model performance.

Communication efficiency. Fig.4 shows the average test accuracy of all clients with different number of communication rounds. With a small number of rounds (*e.g.*, less than 50 on the Epic-Kitchens), our model has similar performance as the baselines, *e.g.*, PerAvg, FedProto, and FedLAW. Thanks to the Hyper-graph Aggregation, our model achieves consistently better accuracy than all baselines after more rounds of training. As demonstrated in Fig.4a, the



(a) EPIC-Kitchens (b) UCF-101 (c) MELD

Figure 4. Convergence rate of each method on three datasets.

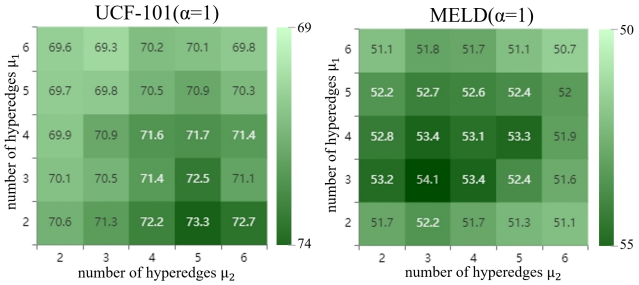


Figure 5. Test accuracy (%) overview for HAMFL with varying hyperedges(μ_1, μ_2).

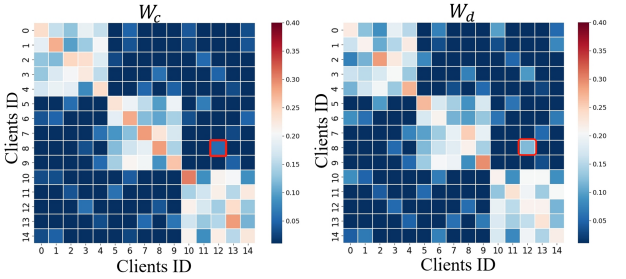


Figure 6. Visualization of weight scores (W_c, W_d) for hypergraph aggregated Modality-Specific(Left) and Modality-Shared(Right) modules.

performance of HAMFL is rapidly boosted as soon as the hypergraph model on the server converges.

Varying Number of hyperedges (μ_1, μ_2). We use μ_1 and μ_2 to denote the number of hyperedges in the Modality Speculative Domain and the Distributed Speculative Domain, respectively. We train with different numbers of hyperedges; specifically, we set μ_1, $\mu_2 \in \{2, 3, 4, 5, 6\}$. Fig.5 shows the results on UCF-101 ($m = 2$) and MELD ($m = 3$) datasets. For μ_1, optimal performance is achieved when it aligns with the number of modalities in the dataset (i.e., $\mu_1 = m$). It indicates that HAMFL can accurately identify features of different modalities, effectively clustering clients with similar modalities. The optimal values of μ_2 are 5 (for UCF-101) and 3 (for MELD), respectively. The UCF dataset requires larger μ_2 due to the increased diversity in data distribution caused by its wider variety of labels. Such statistical characteristics demand more refined clustering to ensure that clients with similar data distributions are effectively clustered.

Visualization of Multimodal Hypergraph Aggregation. Fig.6 illustrates the visualized parameter aggregation weight matrices for Modality-Specific (W_c) and Modality-Shared (W_d) modules. Taking the MELD dataset as an example, we distribute the data of each modality across five clients, amounting to a total of 15 clients. It is clearly observable from the matrices that clients within the same modality exhibit higher weight scores, indicating that the proposed Modality Speculative Domain can accurately deduce the modality of client data. Furthermore, compared with the Modality-Specific module, the Modality-Shared module is more adept at focusing on modality-independent knowledge, such as data distributions. As a result, clients with similar data distributions are given higher weight scores in the aggregation process of W_d, as shown by the red boxes in Fig.6.

5. Conclusion & Limitations

In this paper, we embark on an exploratory journey into modality-agnostic federated learning. Our primary contribution is the Adaptive Hyper-graph Learning Strategy, serving as the cornerstone of our multi-modal client aggregation process on the server side. Additionally, the Global Modality-Shared Prototype has been crucial in assimilating and broadcasting global consensus knowledge within the network. Extensive experimental validation underscores the efficacy of our HAMFL. We acknowledge limitations in data availability and the need for optimization, as well as challenges in evaluating the generalization of federated learning models. Addressing these issues is a priority for our future work to enhance the field.

Acknowledgements. This work was supported by NSFC (No.62206200, 62206137, 62036012, 62376196, U23A20387), and Tianjin Natural Science Foundation (No.22JCQNJC00940, 22JCYBJC00030).

References

[1] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021. 1

[2] Bless Lord Y Agbley, Jianping Li, Amin Ul Haq, Edem Kwedzo Bankas, Sultan Ahmad, Isaac Osei Agyemang, Delanyo Kulevome, Waldiodio David Ndiaye, Bernard Cobbinah, and Shoistamo Latipova. Multimodal melanoma detection with federated learning. In *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 238–244. IEEE, 2021. 2

[3] Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021. 3

[4] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016. 4

[5] Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020. 1

[6] Jiayi Chen and Aidong Zhang. Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 87–96, 2022. 2, 3, 6, 7

[7] Sijia Chen and Baochun Li. Towards optimal multi-modal federated learning on non-iid data with hierarchical gradient blending. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1469–1478. IEEE, 2022. 2

[8] Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are allset: A multiset function framework for hypergraph neural networks. *arXiv preprint arXiv:2106.13264*, 2021. 3

[9] Yutong Dai, Zeyuan Chen, Junnan Li, Shelby Heinecke, Lichao Sun, and Ran Xu. Tackling data heterogeneity in federated learning with class prototypes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7314–7322, 2023. 1

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 6

[11] Jian-hui Duan, Wenzhong Li, Derun Zou, Ruichen Li, and Sanglu Lu. Federated learning with data-agnostic distribution fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8074–8083, 2023. 1

[12] Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. Fedmultimodal: A benchmark for multimodal federated learning. *arXiv preprint arXiv:2306.09486*, 2023. 2

[13] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3558–3565, 2019. 3

[14] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10112–10121, 2022. 1

[15] Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019. 1

[16] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020. 1

[17] Chaoyang He, Emir Ceyani, Keshav Balasubramanian, Murali Annavaram, and Salman Avestimehr. Spreadgnn: Serverless multi-task federated learning for graph neural networks. *arXiv preprint arXiv:2106.02743*, 2021. 1

[18] Jing Huang and Jie Yang. Unignn: a unified framework for graph and hypergraph neural networks. *arXiv preprint arXiv:2105.00956*, 2021. 3

[19] Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao. Dynamic hypergraph neural networks. In *IJCAI*, pages 2635–2641, 2019. 3

[20] Qinbin Li, Bingsheng He, and Dawn Song. Practical one-shot federated learning for cross-silo setting. *arXiv preprint arXiv:2010.01017*, 2020. 6

[21] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021. 1

[22] Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. Revisiting weighted aggregation in federated learning with neural networks. In *International Conference on Machine Learning*, pages 19767–19788. PMLR, 2023. 6, 7

[23] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33:2351–2363, 2020. 6

[24] Yi-Ming Lin, Yuan Gao, Mao-Guo Gong, Si-Jia Zhang, Yuan-Qiao Zhang, and Zhi-Yuan Li. Federated learning on multimodal data: A comprehensive survey. *Machine Intelligence Research*, pages 1–15, 2023. 2

[25] Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *CoRR*, abs/2002.10619, 2020. 1

[26] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 6, 7

[27] Chuizheng Meng, Sirisha Rambhatla, and Yan Liu. Cross-node federated graph neural network for spatio-temporal data modeling. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 1202–1211, 2021. 1

[28] Mahdi Morafah, Saeed Vahidian, Weijia Wang, and Bill Lin. Flis: Clustered federated learning via inference similarity for non-iid data distribution. *IEEE Open Journal of the Computer Society*, 4:109–120, 2023. 1

[29] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018. 6

[30] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019. 1

[31] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32 (8):3710–3722, 2020. 1

[32] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017. 1

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[34] Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8432–8440, 2022. 6, 7

[35] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. *arXiv preprint arXiv:2002.06440*, 2020. 6

[36] Peihao Wang, Shenghao Yang, Yunyu Liu, Zhangyang Wang, and Pan Li. Equivariant hypergraph diffusion neural operators. In *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 5

[37] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1828–1838, 2020. 2

[38] Nannan Wu, Li Yu, Xuefeng Jiang, Kwang-Ting Cheng, and Zengqiang Yan. Fednoro: Towards noise-robust federated learning by addressing class imbalance and label noise heterogeneity. *arXiv preprint arXiv:2305.05230*, 2023. 1

[39] Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. *arXiv preprint arXiv:2306.11867*, 2023. 1

[40] Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3063–3071, 2022. 2, 3, 5, 6, 7

[41] Rui Ye, Zhenyang Ni, Fangzhao Wu, Siheng Chen, and Yanfeng Wang. Personalized federated learning with inferred collaboration graphs. 2023. 1, 3, 6, 7

[42] Rui Ye, Mingkai Xu, Jianyu Wang, Chenxin Xu, Siheng Chen, and Yanfeng Wang. Feddisco: Federated learning with discrepancy-aware collaboration. *arXiv preprint arXiv:2305.19229*, 2023. 1

[43] Jaehyuk Yi and Jinkyoo Park. Hypergraph convolutional recurrent neural network. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3366–3376, 2020. 3

[44] Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2

[45] Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. *arXiv preprint arXiv:2302.08888*, 2023. 6, 7

[46] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pages 7252–7261. PMLR, 2019. 6

[47] Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pages 26311–26329. PMLR, 2022. 6

[48] Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020. 1

[49] Rongyu Zhang, Xiaowei Chi, Guiliang Liu, Wenyi Zhang, Yuan Du, and Fangxin Wang. Unimodal training-multimodal prediction: Cross-modal federated learning with hierarchical aggregation. *arXiv preprint arXiv:2303.15486*, 2023. 2, 5

[50] Yuchen Zhao, Payam Barnaghi, and Hamed Haddadi. Multimodal federated learning on iot data. In *2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 43–54. IEEE, 2022. 2, 5

[51] Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pages 12878–12889. PMLR, 2021. 1

[52] Linlin Zong, Qiujie Xie, Jiahui Zhou, Peiran Wu, Xianchao Zhang, and Bo Xu. Fedcmr: Federated cross-modal retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1672–1676, 2021. 2