Multimodal Fusion Using Multi-View Domains for Data Heterogeneity in Federated Learning

Min Gao[1], **Haifeng Zheng**[1*], **Xinxin Feng**[1], **Ran Tao**[2]

[1]College of Physics and Information Engineering, Fuzhou University, Fuzhou, China
[2]School of Information and Electronics, Beijing Institute of Technology, Beijing, China
221110027@fzu.edu.cn, zhenghf@fzu.edu.cn, fxx1116@fzu.edu.cn, rantao@bit.edu.cn

Abstract

Multimodal information plays an important role in the advanced Internet of Things (IoT) in the era of 6G, which provides reliable and comprehensive assistance for downstream tasks through further fusion and analysis via federated learning (FL). One of the primary challenges in FL is data heterogeneity, which may lead to domain shifts and sharply different local long-tailed category distribution across nodes. These issues hinder the large-scale deployment of FL in IoT applications equipped with multiple various multimodal sensors due to performance deterioration. In this paper, we propose a novel multimodal fusion framework to tackle the aforementioned coupled problems arising during the cooperative fusion of multimodal information without privacy exposure among decentralized nodes equipped with diverse sensors. Specifically, we introduce a flexible global logit alignment (GLA) method based on multi-view domains. This method enables the fusion of diverse multimodal information with the consideration of domain shifts caused by modality-based data heterogeneity. Furthermore, we propose a novel local angular margin (LAM) scheme, which dynamically adjusts decision boundaries for locally seen categories while preserving global decision boundaries for unseen categories. This effectively mitigates severe model divergence caused by significantly different category distributions. Extensive simulations demonstrate the superiority of the proposed framework, which exhibits significant merits in tackling model degeneration caused by data heterogeneity and enhancing modality-based generalization for heterogeneous scenarios.

Introduction

The next-generation wireless networks are driving the rapid development of a series of advanced Internet of Things (IoT) services, such as smart cities, intelligent factories, autonomous driving, and so on (Li et al. 2023a; Zheng et al. 2021). Toward this vision, numerous heterogeneous IoT nodes equipped with various modal sensors are capable of sensing information from multiple domains and centralize the multimodal data at the cloud server for fusion and analysis to aid downstream tasks. However, directly sharing information severely violates privacy protection principles, as the shared data may contain the privacy details of users such as behavioral preferences, geographical locations, and so on.

Federated learning (FL), as a typical cooperative training paradigm, provides an effective solution for privacy protection by delivering the updated model learned from private data to the server instead of directly uploading the private data (McMahan et al. 2017). Accordingly, FL-oriented works toward multimodal fusion have attracted more and more attention. Most existing works focus on improving model performance by incorporating the complementary information from multiple sensors embedding on decentralized nodes where the multimodal fusion is extended from the traditional centralized operations (Feng et al. 2023b; Khoa et al. 2024). These methods do not consider the data heterogeneity especially in modality-based heterogeneity in practical multimodal FL scenarios. Although several FL works for data heterogeneity consider either long-tail category distribution (Li et al. 2023b; Xiao et al. 2024) or domain shifts (Feng et al. 2023a; Chowdhury et al. 2022) between nodes, most of them do not consider the aforementioned issues in the context of coupled modality-based and category-based data heterogeneity in multimodal FL scenarios. These issues become more complex in multimodal FL scenarios for the introduction of modality-based heterogeneity between nodes, which leads to significantly degraded performance caused by worsening model divergence.

Specifically, domain shifts (same label, different features) is widespread in real-world FL applications. For example, different cameras/scanners and imaging protocols of different nodes present heterogeneity, which leads to domain shift within a single modality (e.g., image) of data (Feng et al. 2023a). However, the domain shifts in the heterogeneous multimodal FL scenarios become more complex, which involves from solely intra-modality level to coupled intra-modality and inter-modality level. To the best of our knowledge, there is no previous FL works focusing on this complex domain shifts in heterogeneous multimodal FL scenarios. Furthermore, long-tailed distribution inevitably exists due to the fixed geographical locations or the behavioral preferences, which forms non-$i.i.d.$ (not independently and identically distributed) data among nodes and would enables the global data manifests a long-tailed distribution. Existing works only focus on the category-wise imbalance (Zeng et al. 2023; Li et al. 2023b; Xiao et al. 2024), but overlook

*Corresponding Author.

the modality-wise imbalance arisen in the context of multimodal heterogeneous FL scenarios. Actually, it is evident that the samples within a single category may still be long-tailed due to the varying complementary attributes from various modalities of data. This poses challenges to multimodal FL applications that are markedly different from those issues at category-based imbalance level. So far, there has been little attention to those more complex issues caused by coupled modality-based and category-based data heterogeneity in multimodal FL scenarios.

In this paper, we propose a novel multimodal fusion framework for FL with multi-view domain, named FedMVD, to comprehensively consider the aforementioned problems. Unlike previous methods, we define each arbitrary modality or modality set as a view domain of data. Specifically, we design a flexible global logit alignment (GLA) scheme that efficiently splits and projects various data view domains into tailored spaces, which mitigates model divergence arising from intricate domain shifts in multimodal FL scenarios. This is achieved by constructing multi-view domains within a shared model, and then adaptively activating as well as integrating relevant domains during local training according to modalities owned by nodes. This not only effectively enhances the model reliability for interfacing with data from arbitrary view domains, but also boosts generalization for some unseen view domains by integrating the feature capture capabilities from the related modalities. Additionally, we propose a local angular margin (LAM) scheme to mitigate severe model divergence, which arises from the sharply varying local long-tailed distributions among heterogeneous nodes. This scheme achieves local adjustment based on global decision boundaries, which follows the prior of global long-tailed distribution. Moreover, it dynamically adjusts local decision boundaries among seen categories while maintaining global decision boundaries for unseen categories according to local data distribution, resulting in a refined global consensus through aggregation without exposing privacy such as data distribution to others.

Generally, the main contributions are summarized as follows.

- We propose a novel FedMVD framework to address the more complex domain shifts and long-tailed distribution in multimodal FL scenarios. To the best of our knowledge, this is the first work to address the multimodal FL task with heterogeneous data from multi-view domain perspective.
- We propose an effective GLA scheme to tackle the model divergence caused by domain shifts. GLA has the capability to adaptively activate and integrate relevant domains, which can achieve excellent generalization performance across unseen view domains.
- We propose a novel LAM scheme for more complex long-tailed distributions in multimodal FL scenarios. This scheme dynamically adjusts decision boundaries among seen categories while maintaining global decision boundaries for unseen categories based on local data distribution, which effectively prevents the risk of privacy exposure.
- We evaluate the proposed framework using two real-world multimodal datasets. The results demonstrate the superiority of our framework, which highlights its significant merits in tackling model degeneration caused by coupled modality-based and category-based data heterogeneity and enhancing generalization for unseen view domains in heterogeneous multimodal FL scenarios.

Related Works

Multimodal Federated Learning

FL-oriented works toward multimodal fusion have attracted more and more attention due to the model performance improvements which benefit from the complementary information provided by various decentralized modal sensors (Feng et al. 2023b; Khoa et al. 2024). However, the multimodal fusion in these works are extended from the traditional centralized operations, such as concatenation (Poria et al. 2016), attention-based fusion (Liu et al. 2023), and tensor fusion (Xueming Yan and Liu 2022; Ben-younes et al. 2017), without the consideration of coupled modality-based and category-based data heterogeneity. In this paper, we design a novel multimodal fusion framework based on multi-view domains to tackle the model degeneration caused by coupled modality-based and category-based data heterogeneity in multimodal FL scenarios.

Federated Learning for Data Heterogeneity

Long-Tailed Learning. Long-tailed distribution ubiquitously exists in real-world data at large-scale applications, where head categories have the majority of samples while tail categories have only a minority (Tao et al. 2023). Numerous efforts have been undertaken to tackle the data imbalance problem, such as model decouple for classifier rebalance (Li et al. 2023b; Xiao et al. 2024; Dai et al. 2023; Tan et al. 2022; Zhang et al. 2024), cost-sensitive learning (Zeng et al. 2023; Shi et al. 2024), calibration (Ye et al. 2023; Zhang et al. 2024; Dai et al. 2024), aggregation (Tan et al. 2022; Hu et al. 2024), and so on. However, existing works towards category-wise without considering modality-wise imbalance are not comprehensive, because the samples within a category may be long-tailed due to the varying complementary attributes from various modalities of data.

Domain Shifts. Domain shifts is one of the prevalent data heterogeneous property in FL because data from independently heterogeneous nodes often form distinct source domains (Huang et al. 2023; Zhang et al. 2023). Existing works have focused on addressing the poor generalization caused by domain shifts. The main techniques contains prototype learning (Huang et al. 2023), alignment for source domain (Zhang et al. 2023), aggregation (Jiang, Wang, and Dou 2022; Shenaj et al. 2023), and so on. However, existing intra-modality works are not comprehensive for heterogeneous multimodal FL scenarios due to more complex domain shifts involving coupled inter-modality and intra-modality.

In this paper, we conduct an in-depth analysis of the aforementioned complex problems that arise in multimodal FL scenarios due to coupled modality-based and category-based data heterogeneity and propose solutions.

System Model and Problem Statement

System Model

We consider a range of widely used applications in which multiple decentralized nodes equipped with heterogeneous sensors cooperate to improve model performance for more comprehensive and powerful capabilities.

The aforementioned scenario can be simplify as shown in Fig. 1, which involves N nodes $\{E_1, \ldots, E_N\}$ and a cloud server S. There are a total of C categories in the label space and M different data modalities in the multimodal FL framework. For an arbitrary node $E_i, i \in [N]$, its private dataset with both the available modalities $\mathcal{M}_i \subseteq \{M_1, M_2, \ldots, M_M\}$ and categories set C_i is denoted as $D_i : \{(\{x_j^k | k \in \mathcal{M}_i\}, y_j)\}_{j=1}^{|D_i|}$, where x_j^k is the k-th modality of the j-th training sample and y_j is the corresponding label whose categorical distribution is $p_i(y)$. Consequently, the global training dataset which is the centralized private data from all nodes is denoted as $\mathcal{D}_{tr} = \bigcup_{i=1}^{N} D_i$, and its categorical distribution is $p_{tr}(y) : \sum_{i=1}^{N} |D_i| p_i / \sum_{i=1}^{N} |D_i|$. The testing dataset is denoted as $\mathcal{D}_{te} = \{(\{x_j^k | k \in \mathcal{M}_s, \mathcal{M}_s \in \mathcal{M}_{te}\}, y_j)\}_{j=1}^{|\mathcal{D}_{te}|}$, where $\mathcal{M}_{te} = \{\mathcal{M}_s \subseteq \{M_1, M_2, \ldots, M_M\} | \mathcal{M}_s \neq \emptyset\}$, and the corresponding categorical distribution is $p_{te}(y)$. It is acknowledged that the model would perform well when training and testing data are $i.i.d.$; however, it seldomly holds for real-world environments (Gu et al. 2022), especially in heterogeneous multimodal FL frameworks. The goal of such frameworks is to cooperatively learn a model ω across diverse nodes with local data to achieve robust prediction with arbitrary modalities of data.

Motivation and Problem Statement

The multimodal model $\Psi(\cdot)$ can be divided into the a feature extraction module $\mathcal{I}(\cdot)$, a feature fusion module $\mathcal{J}(\cdot)$, and a feature prediction module $\mathcal{P}(\cdot)$. The estimated error ϵ on testing dataset can be described as:

$$\begin{aligned}\epsilon = \mathbb{E}_{\substack{m \sim P_{\mathcal{D}_{te}} \\ n \sim P_{\mathcal{D}_{tr}} \\ j \sim P_{D_i}}} & \mathbb{E}_{(\boldsymbol{x}, y) \sim p_i} f(\Psi(\boldsymbol{x}), y) \\ & \cdot \underbrace{\frac{p_{te}(y) p_{te}(y|\Psi(\boldsymbol{x}))}{p_{tr}(y) p_{tr}(y|\Psi(\boldsymbol{x}))}}_{\text{global testing and training distribution bias}} \cdot \underbrace{\frac{p_{tr}(y) p_{tr}(y|\Psi(\boldsymbol{x}))}{p_i(y) p_i(y|\Psi(\boldsymbol{x}))}}_{\text{global and local training distribution bias}}, \end{aligned} \tag{1}$$

where $P_{\mathcal{D}_{te}}, P_{\mathcal{D}_{tr}}$, and P_{D_i} denote the probability of sample data m, n, and j from testing dataset, training dataset, and local private dataset, respectively. $f(\cdot)$ is the loss function. According to the Bayes theorem, $p(y|\Psi(\boldsymbol{x}))$ in Eq. (1) can be further converted into $p(y|(z_c, (z_a|\mathcal{M}_i)))$ due to the multimodal information $\boldsymbol{x} = \{x^k | k \in \mathcal{M}_i\}$ can be further described as the class-specific invariant information z_c and the modality-related complementary information z_a (Zuo et al. 2023):

$$p\left(y|(z_c, (z_a|\mathcal{M}_i))\right) = \frac{p\left(z_c|y\right)}{p\left(z_c\right)} \cdot \underbrace{\frac{p\left(z_a|\mathcal{M}_i, y, z_c\right)}{p\left(z_a|\mathcal{M}_i, z_c\right)}}_{\text{modality bias}} \cdot p(y). \tag{2}$$

It is obvious from Eq. (1) and Eq. (2), the coupled category distribution bias and modality bias worse the estimated error ϵ, which leads to severe model deterioration. Hence, to minimize ϵ, there are several challenges to be resolved:

(1) For real-world scenarios, different modalities and viewpoints would alter the distribution and lead to domain shifts. These domain shifts raise the requirements to align the distribution to a global unbiased and semantically meaningful space with diverse $p_i(y|(z_c, (z_a|\mathcal{M}_i))), i \in [N]$.

(2) Due to the preference of behavior patterns or fixed geographical location, there may be some diverse unseen or rare categories at various nodes. Thus, it yields the ill-conditioned decision boundary for tail categories since the local tail classifier is over-suppressed by the head categories. Consequently, sharply different local long-tailed category distribution worse the divergence of local models, which poses the significant challenge of global decision alignment.

To address aforementioned challenges, we propose a novel heterogeneous multimodal FL framework. The details are presented as follows.

The Proposed FedMVD Framework

In this section, the proposed heterogeneous multimodal FL framework (see Fig. 1), which involves global model initialization, local model update, and aggregation stages, is introduced to achieve multiple heterogeneous nodes collaboratively training a shared global model without exposing its private data to each other in multimodal systems. To address both domain shifts and model divergence that come from coupled modality and data heterogeneity, we propose a training paradigm under the FL framework, named FedMVD, which consists of a global logit alignments scheme GLA and a local logit adaptive adjustment scheme LAM.

The Proposed GLA Method

To align various $p_i(y|\Psi(\boldsymbol{x}))$ to an global unbiased and semantically meaningful space with diverse $p_i(y|(z_c, (z_a|\mathcal{M}_i)))$, $i \in [N]$, a global logit alignments scheme is embedded in the proposed framework as shown in Fig. 1. As aforementioned, the multimodal model $\Psi(\cdot)$ can be divided into the a feature extraction module $\mathcal{I}(\cdot)$, a feature fusion module $\mathcal{J}(\cdot)$, and a feature prediction module $\mathcal{P}(\cdot)$. Note that, this paper solely focuses on addressing the issues of both domain shifts and model divergence during the process of multimodal data fusion among FL framework to make full use of information from various heterogeneous nodes which may equip with different types/numbers of sensors. Consequently, the designs of both the feature extraction module $\mathcal{I}(\cdot)$ and the feature prediction module $\mathcal{P}(\cdot)$ are not considered in this paper.

Given the output of M modal feature extraction submodules are $\{\mathcal{Z}^{(n)} | n \in [M]\}$ where $\mathcal{Z}^{(n)} \in \mathbb{R}^{I_n}$, and the fusion

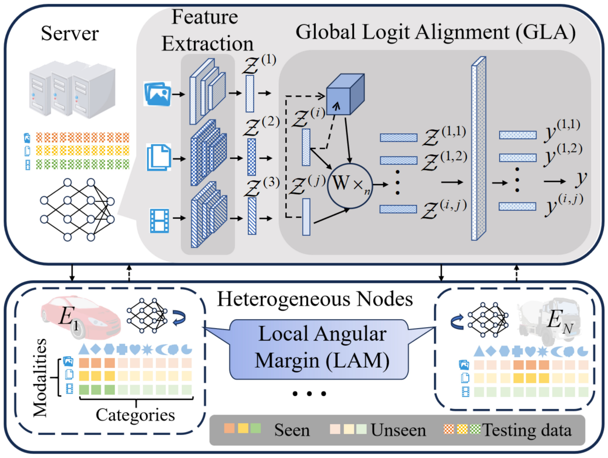


Figure 1: The overview of the proposed FedMVD framework.

tensor is $\mathcal{T} \in \mathbb{R}^{I_1 \times \cdots \times I_M \times |A|}$, the output of multilinear feature fusion module can be represented as follows:

$$\mathcal{J}(\{\mathcal{Z}^{(n)} | n \in [M]\}) = \mathcal{T} \times_1 \mathcal{Z}^{(1)} \cdots \times_M \mathcal{Z}^{(M)}. \quad (3)$$

Correspondingly, the j-th element of output can be described as follows:

$$\sum_{i_M}^{I_M} \left(\cdots \sum_{i_2}^{I_2} \left(\sum_{i_1}^{I_1} \mathcal{T}_{i_1, i_2, \cdots, i_M, j} \cdot \mathcal{Z}_{i_1}^{(1)} \right) \cdot \mathcal{Z}_{i_2}^{(2)} \cdots \right) \cdot \mathcal{Z}_{i_M}^{(M)}. \quad (4)$$

MUTAN (Ben-younes et al. 2017) is a typical multimodal tensor-based Tucker decomposition to efficiently address huge dimensionality issues of fusion tensor $\mathcal{T}$ in Eq. (3). However, it is inappropriate to directly apply the traditional centralized multimodal learning models such as MUTAN as the shared model in FL for heterogeneous scenarios. This is because, as shown in Eq. (4), all modalities are evidently projected into the same joint-all-modalities space where each element solely depends on the all modality data for which multimodal fusion model design. In practice, it is cannot guarantee that all modalities are always present at all nodes in heterogeneous scenarios, which leads to significant unaligned joint-all-modalities space among heterogeneous nodes equipped with different types/numbers of sensors.

Motivated by these issues, we design a flexible GLA scheme for domain shifts. This scheme constructs multi-view domains in the shared global model to simulate various modality heterogeneity. As shown in Fig. 1, a multi-view fusion module $\Phi(\cdot)$ is constructed to create a joint feature representation space and capture the correlations and interactions between modalities in the same space. For a multimodal system which processes M modalities, the multi-view domain of the joint feature representation between the i-th and j-th modalities is constructed as follows:

$$\begin{aligned} \mathcal{Z}^{(i,j)} &= \Phi(\mathcal{Z}^{(i)}, \mathcal{Z}^{(j)}) \\ &= ((\mathcal{T}_c \times_1 \underbrace{((\mathcal{Z}^{(i)})^\top U_1)}_{\tilde{\mathcal{Z}^{(i)}} \in \mathbb{R}^{r_1}}) \times_2 \underbrace{((\mathcal{Z}^{(j)})^\top U_2)}_{\tilde{\mathcal{Z}^{(j)}} \in \mathbb{R}^{r_2}}) \times_3 U_3, \end{aligned} \quad (5)$$

where $\mathcal{T}_c \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, $\{U_s \in \mathbb{R}^{I \times r_s} | 0 < s \leq 3\}$, and each feature extraction dimension is I. According to the tensor sparsity theory (Ben-younes et al. 2017), the model parameters can be reduced by representing $\mathcal{T}_c \times_1 \tilde{\mathcal{Z}^{(i)}} \times_2 \tilde{\mathcal{Z}^{(j)}}$ as $\sum_{r=1}^{R} (\tilde{\mathcal{Z}^{(i)}}^\top M_r) * (\tilde{\mathcal{Z}^{(j)}}^\top N_r)$ with both R matrices $M_r \in \mathbb{R}^{r_1 \times r_3}$ and R matrices $N_r \in \mathbb{R}^{r_2 \times r_3}$, where $*$ denotes element-wise multiplication.

Then, the output of the multi-view fusion can be described as follows:

$$\begin{aligned} \mathcal{J}(\{\mathcal{Z}^{(n)} | n \in [M]\}) &= \{\Phi(\mathcal{Z}^{(i)}, \mathcal{Z}^{(j)}) | i \in [M], j \in [M]\} \\ &= \{\mathcal{Z}^{(i,j)} | i \in [M], j \in [M]\}, \end{aligned} \quad (6)$$

where $\mathcal{Z}^{(i,j)}$ contains invariant information $z_c^{(i,j)}$ and the complementary information $z_a^{(i,j)}$, i.e.,$\mathcal{Z}^{(i,j)} = [z_c^{(i,j)}, z_a^{(i,j)}]$. In this way, the interaction in any modality-based view between the i-th and j-th modalities are captured in the same way which all correlations $\tilde{\mathcal{Z}^{(i)}}[m]\tilde{\mathcal{Z}^{(j)}}[n]$ are projected to the same joint representation space. As a result, the distribution of $p_i(y|\Psi(\boldsymbol{x}))$ achieves aligning to an global unbiased and semantically meaningful space with diverse $p_i(y|(z_c, (z_a|\mathcal{M}_i))), i \in [N]$.

Finally, the prediction is obtained based on the average logits of multi-view domains to integrate the knowledge of different modality-based views as follows:

$$\begin{aligned} y &= \mathcal{P}(\mathcal{J}(\{\mathcal{Z}^{(n)} | n \in [M]\})) \\ &= \frac{1}{|\mathcal{Z}^{(i,j)}|} \sum_{i,j}^{M} \mathcal{P}(\underbrace{\Phi(\mathcal{Z}^{(i)}, \mathcal{Z}^{(j)})}_{[z_c^{(i,j)}, z_a^{(i,j)}]}). \end{aligned} \quad (7)$$

In this way, the nonexistent modalities enable a part of the correlated view ineffective while making no difference to the relative size of prediction elements due to multi-view-based average operation. It means that the proposed method enables aware the more importance invariant information, which exists in more numbers of modalities, for each class. While the complementary information which depends on the modality and data heterogeneity across decentralized nodes provides the supplementary for an object without affecting the aware of the invariant information. As a result, the proposed method enables the multimodal nodes learn more compact invariant information feature representations than single modal nodes while mitigates model divergence due to modality heterogeneity across nodes.

The Proposed LAM Method

The typical softmax cross-entropy loss suffers from performance degradation due to the ill-conditioned decision caused by long-tailed data distribution (Zeng et al. 2023; Tan et al. 2021). This problem will be worse under heterogeneous multimodal FL scenario when the distribution of local data is extremely non-$i.i.d.$ leading to misaligned classification space.

To address aforementioned problems, we propose a novel LAM method based on cosine distance-based softmax loss to adaptive adjust decision boundary among local long-tailed setting while keeping decision boundaries of unseen

Dataset	Scenarios		Methods	Modality-Based Heterogeneous Setting								
				Θ ={0,0,1}			Θ ={0,0.6,0.4}			Θ ={0.6,0,0.4}		
				mAccu	mPrec	mF1	mAccu	mPrec	mF1	mAccu	mPrec	mF1
FLASH	LOS		FedGPI	70.46	59.56	57.31	61.30	45.55	41.10	51.88	40.56	36.07
			FedFM	51.88	30.04	27.29	47.73	28.14	24.90	34.31	18.63	16.49
			FedETF	59.53	57.15	48.98	54.34	52.26	44.19	33.04	32.60	22.70
			FedLoGe	55.70	29.46	25.78	49.36	25.79	21.63	45.24	23.99	19.87
			Ours	**72.27**	**65.01**	**61.10**	**65.01**	**55.06**	**51.71**	**52.36**	**42.85**	**38.40**
	NLOS	Pedestrian	FedGPI	70.56	63.30	58.58	62.48	49.06	43.00	55.18	45.01	40.66
			FedFM	55.82	30.51	26.19	47.89	26.18	22.55	46.39	24.44	23.58
			FedETF	67.68	60.00	58.80	53.42	48.96	46.02	34.19	35.67	31.89
			FedLoGe	48.48	22.44	15.52	44.59	15.96	9.36	46.99	15.92	14.02
			Ours	**75.68**	**64.02**	**59.81**	**66.40**	**54.18**	**49.04**	**56.75**	**49.07**	**46.83**
		Moving Car	FedGPI	69.67	60.21	56.21	62.73	47.26	42.80	**54.52**	32.86	29.78
			FedFM	60.23	26.05	23.61	55.75	23.04	21.52	39.72	14.85	14.35
			FedETF	56.64	48.36	46.71	48.38	44.93	40.66	38.99	26.09	21.71
			FedLoGe	62.83	31.58	26.43	53.59	21.46	17.51	48.57	16.14	13.27
			Ours	**72.57**	**65.99**	**64.61**	**67.90**	**49.82**	**47.12**	52.31	**35.00**	**31.10**
UTD-MHAD	$Dir(0.1)$		FedGPI	93.21	93.65	92.60	84.98	84.82	82.64	80.25	82.23	77.95
			FedFM	80.66	85.14	80.59	76.75	75.97	72.95	66.39	61.63	60.41
			FedETF	78.81	79.48	76.64	67.28	70.22	65.06	63.92	66.83	60.86
			FedLoGe	89.85	90.75	88.87	84.36	83.97	81.91	74.28	75.47	71.02
			Ours	**95.54**	**95.79**	**95.37**	**87.38**	**87.73**	**85.40**	**82.51**	**84.68**	**79.95**

Table 1: Performance comparisons with FL-oriented methods for data heterogeneity. The first and second highest scores are represented by **bold** font and underline, respectively.

categories. Specifically, we add a category-aware adjustment function $M(\cdot)$ to encourage inter-class discrimination and enhance the intra-class compactness by adjusting the decision boundaries between the ground truth categories i and all the other non ground truth categories j at the k-th node:

$$L_s = -\log \frac{e^{s \cos \theta_i}}{\sum_{j=1}^{C} e^{s \cos(\theta_j + M(i,j))}}, \tag{8}$$

where

$$M\left(i,j\right) = \frac{\alpha}{\pi} \log(\frac{n_i + \varepsilon}{n_j + \varepsilon}). \tag{9}$$

n_j is the number of category j at the k-th node, ε and α are hyper-parameters to prevent the denominator from zero and control the strength of pushing the decision boundary, respectively. They are set as small values in our experiments.

The proposed method meticulously explores subcategories representation ability of all node and eliminates ambiguity from various unseen categories at different nodes. Specifically, when both the two categories i and j belong to the k-th node, the decision boundary is pushed towards the classifier weight vector with smaller class cardinalities, that is the decision boundary will move towards category j when $n_i > n_j$ or category i when $n_i < n_j$. When either the two categories y_i or j not belong to the k-th node, the proposed loss degrades to the softmax loss with cosine distance, since the k-th node does not have samples to support decision boundary adjustments. When one category belongs to the k-th node while the others not, the decision boundary will move towards existing category without ill-conditioned decision boundary. This is because the $\log(\cdot)$ function will output a value smaller than the input, and the output will increase slower as the input becomes larger because its second derivative is smaller than 0 (Wang et al. 2022).

Performance Evaluation

We evaluate the proposed methods on two typical multi-modal datasets named FLASH (Salehi et al. 2022; Ouyang et al. 2023) and UTD-MHAD (Chen, Jafari, and Kehtarnavaz 2015; Ouyang et al. 2022), and conduct ablation study on UTD-MHAD dataset.

Evaluation Metrics

We report the mean accuracy (mAccu), precision (mPrec), and F1-score (mF1) of the best single global model for all view domains (all combinations of modalities) of the testing dataset to evaluate both performance and generalization.

Implementation Details

To evaluate the effectiveness of our framework for heterogeneous scenarios, we simulate modality-based data heterogeneity on the naturally heterogeneous FLASH dataset containing 210 heterogeneous nodes, and both categories-based and modality-based data heterogeneity on the UTD-MHAD dataset.

Dataset	Scenarios		Methods	Modality-Based Heterogeneous Setting								
				Θ ={0,0,1}			Θ ={0,0.6,0.4}			Θ ={0.6,0,0.4}		
				mAccu	mPrec	mF1	mAccu	mPrec	mF1	mAccu	mPrec	mF1
FLASH	LOS		FedCon	56.68	50.07	46.72	46.73	36.03	34.45	37.96	31.93	28.44
			FedAtt	59.32	49.65	46.16	45.78	43.48	40.78	38.34	32.20	27.69
			MUTAN	51.61	49.97	47.28	52.03	45.61	41.00	32.27	23.99	20.91
			Ours	**72.27**	**65.01**	**61.10**	**65.01**	**55.06**	**51.71**	**52.36**	**42.85**	**38.40**
	NLOS	Pedestrian	FedCon	58.89	54.80	49.36	53.50	44.36	41.45	40.97	37.64	34.21
			FedAtt	61.24	39.54	39.23	45.90	35.10	31.47	34.27	21.70	21.72
			MUTAN	60.46	49.46	44.57	57.54	39.65	37.52	44.11	28.18	28.46
			Ours	**75.68**	**64.02**	**59.81**	**66.40**	**54.18**	**49.04**	**56.75**	**49.07**	**46.83**
		Moving Car	FedCon	45.97	41.81	36.15	53.74	37.96	34.92	36.48	23.18	22.41
			FedAtt	65.63	52.76	51.66	53.24	29.09	26.91	41.35	22.87	20.72
			MUTAN	61.31	53.38	52.33	54.87	33.71	31.78	41.64	20.17	19.37
			Ours	**72.57**	**65.99**	**64.61**	**67.90**	**49.82**	**47.12**	**52.31**	**35.00**	**31.10**
UTD-MHAD	$Dir(0.1)$		FedCon	89.03	90.05	88.25	85.87	85.00	83.53	77.37	74.41	73.21
			FedAtt	59.40	57.97	55.01	63.65	60.87	58.86	60.56	57.69	55.12
			MUTAN	79.90	80.73	77.35	67.76	60.94	61.37	66.87	63.83	61.35
			Ours	**95.54**	**95.79**	**95.37**	**87.38**	**87.73**	**85.40**	**82.51**	**84.68**	**79.95**

Table 2: Performance comparisons with FL-oriented multimodal fusion method.

Heterogeneous Settings. We simulate the heterogeneous setting in terms of categories and modalities. For category-based heterogeneous setting, following the previous works (Huang et al. 2021; Li, He, and Song 2021), we generate non-$i.i.d.$ data partition with Dirichlet distribution, denoted as $Dir(\beta)$, where the smaller β controls more heterogeneous setting (0.1 by default). As for the modality-based heterogeneous setting, similar to previous works (Ouyang et al. 2023), we apply $\Theta = \{\alpha_m | m \in M\}$ to control the ratios of nodes possessing m modalities, where $\sum_{m=1}^{M} \alpha_m = 1$.

Compared Methods. We consider two types of FL-oriented baselines to evaluate the proposed FedMVD: (1) Data heterogeneity methods, including FedGPI (Zeng et al. 2023), FedFM (Ye et al. 2023), FedETF (Li et al. 2023b), and FedLoGe (Xiao et al. 2024). (2) Multimodal fusion methods, including FedMultimodal (Feng et al. 2023b) which contains Concatanation-based fusion (FedCon) and Attention-based Fusion (FedAtt), and MUTAN embedded in FL framework (Ben-younes et al. 2017).

Comparisons with Previous Methods

Comparisons with FL-Oriented Methods for Data Heterogeneity. As shown in Table 1, the proposed FedMVD consistently demonstrates remarkable performance and generalization in various heterogeneous settings. For example, in an extremely heterogeneous setting in our experiments ($\Theta = 0.6, 0, 0.4$, where 60% of nodes possess arbitrary single modality data and 40% of nodes have all modality data), FedMVD achieves 82.51%, 84.68%, and 79.95% on the UTD-MHAD dataset in terms of mAccu, mPrec, and mF1, respectively, surpassing the second best method FedGPI (Zeng et al. 2023) by 2.26%, 2.45%, and 2%. Similar improvements are observed in the naturally heterogeneous FLASH dataset. This significant enhancement is attributed to dynamical alignment of decision boundaries across heterogeneous nodes and optimization of local decision boundaries based on local private data. Specifically, comparisons with FedGPI (82.51% vs. 80.25%), which adopts gradient magnitude as prior information in local re-balance loss to mitigate performance degradation from data heterogeneity, highlight the importance of aligning decision boundaries across nodes. Furthermore, the superiority of FedMVD over classifier calibration methods such as FedETF (82.51% vs. 63.92%) and FedLoGe (82.51% vs. 74.28%), which essentially use a fixed simplex equiangular tight frame (ETF) structure to solve classifier bias. It implies that dynamically adjusting decision boundaries based on local private data is more effective for diverse heterogeneous scenarios.

Comparisons with FL-Oriented Multimodal Fusion Methods. As demonstrated in Table 2, the FedMVD consistently outperforms the other typical FL-oriented multimodal fusion methods across all metrics (e.g., 82.51%, 84.68%, and 79.95% in $\Theta = \{0.6, 0, 0.4\}$ on UTD-MHAD). This remarkable performance is due to the consideration of multi-view domain which addresses coupled modality-based and category-based data heterogeneity in complex and various scenarios. Previous works like FedCon, FedAtt (Feng et al. 2023b), and MUTAN (Ben-younes et al. 2017) fuse fixed data modalities (i.e., a certain view domain of data), which limits their effectiveness in heterogeneous environments due to the following issues. Firstly, these works design the multimodal fusion model for inputs with fixed modalities, which cannot always be ensured to be captured in practice. This design may cause the multimodal mod-

GLA	LAM	mAccu	mPrec	mF1
-	-	66.87	63.83	61.35
✓	-	80.32	80.12	77.08
-	✓	68.59	66.16	63.45
✓	✓	82.51	84.68	79.95

Table 3: Ablation study of each component of our framework with the modality-based heterogeneous setting $\Theta = \{0.6, 0, 0.4\}$ and category-based heterogeneous setting non-$i.i.d.$ with $Dir(0.1)$.

el to have specialization in a certain view domain of data but weaknesses in others. Secondly, coupled modality-based and category-based data heterogeneity results in more severe local model divergence. FedMVD overcomes these limitations with its multi-view-based multimodal fusion, which enables effective fusion across multiple heterogeneous nodes.

Ablation Study

Effects of Each Component. The proposed FedMVD involves two novel training paradigms, including the GLA to align the global decision boundary for domain shifts, and the LAM to adaptively adjust the local decision boundary. To investigate the effectiveness of GLA and LAM ($+1.72\%$ for non-$i.i.d$), we conduct ablation studies with the modality-based heterogeneous setting $\Theta = \{0.6, 0, 0.4\}$. As demonstrated in Table 3, GLA delivers mAccu, mPrec, and mF1 improvements of 13.45%, 16.29%, and 15.73% over the baseline for non-$i.i.d.$, which demonstrates the merits of GLA in addressing model degeneration in terms of both performance and generalization caused by complex domain shifts in multimodal FL scenarios. Furthermore, LAM surpasses the baseline by 1.72%, 2.33%, and 2.10% in mAccu, mPrec, and mF1 for non-$i.i.d.$, respectively, which indicates the effectiveness of LAM for negative effect caused by long-tailed distribution.

Beyond Simple Long-Tail Learning Extension. To validate the uniqueness of the proposed LAM scheme for long-tailed distribution in FL scenarios, we conduct experiments with long-tailed learning-oriented methods, including ArcFace (Deng et al. 2019), C2AM (Wang et al. 2022), CosFace (Wang et al. 2018), BSM (Ren et al. 2020), and GCL (Li, Cheung, and Lu 2022), by replacing the proposed LAM scheme embedded in FedMVD with these methods. As shown in Table 4, the proposed method exhibits a significant superiority compared to the other simple long-tail learning extension. This is because the unaligned ill-conditioned decision boundary across non-$i.i.d.$ nodes influences the convergence of the global model. The proposed LAM scheme address this problem by considering two aspects: 1) the decision boundary remains unchanged for unseen categories, or 2) the decision boundary adjusts according to the seen categories based on global knowledge.

Effects of Different Non-$i.i.d.$ Data Distributions. To verify performance of various category-based heterogeneous setting, we conduct experiments with different degree

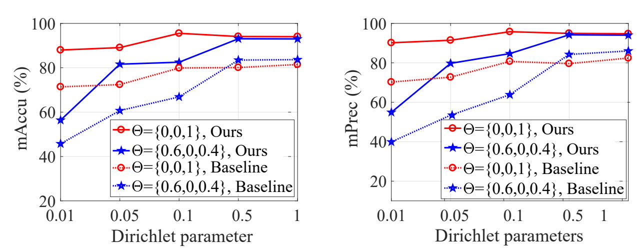
Methods	Modality-Based Heterogeneous Setting					
	Θ ={0,0,1}			Θ ={0.6,0,0.4}		
	mAccu	mPrec	mF1	mAccu	mPrec	mF1
ArcFace	91.29	92.18	90.55	78.44	77.74	75.29
C2AM	91.54	92.32	90.76	78.24	77.70	74.99
CosFace	91.22	92.18	90.49	77.87	77.60	74.75
BSM	90.88	91.53	90.05	77.96	77.70	75.41
GCL	69.62	71.47	67.61	71.33	71.09	67.59
Ours	**94.22**	**94.80**	**93.81**	**81.09**	**82.31**	**78.41**

Table 4: Performance comparisons of long-tail learning-oriented methods.

of heterogeneity by controlling the parameter β in Dirichlet distribution. As shown in Fig. 2, it is evident that the curves of our method demonstrate superiority compared to the baseline in terms of mAccu and mPrec. Furthermore, our method demonstrates a lower sensitivity to β in comparisons to the baseline. For example, our curves become flatter when $\beta \geq 0.05$ than baseline. This reduced sensitivity indicates that our method effectively enhances robustness to variations in the category-based heterogeneous setting.

Conclusion

In this paper, we proposed a novel multimodal fusion framework, termed FedMVD, to address domain shifts and sharply different local long-tailed category distribution caused by the coupled modality-based and category-based data heterogeneity among heterogeneous nodes. The proposed FedMVD incorporates a global alignment scheme GLA based on multi-view domains for domain shifts among heterogeneous nodes, and a local adjustment strategy LAM for severe model divergence caused by significantly different category distributions. Extensive simulations have demonstrated the superiority of the proposed framework, which exhibits significant merits in addressing model degeneration in terms of accuracy, precision, and so on caused by data heterogeneity, and enhancing modality-based generalization for heterogeneous scenarios.



(a) mAccu (b) mPrec

Figure 2: The effects of different heterogeneous scenarios.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 6197113; and in part by the Major Science and Technology Program of Fujian Province under Grant 2022HZ026007.

References

Ben-younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2612–2620.

Chen, C.; Jafari, R.; and Kehtarnavaz, N. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International Conference on Image Processing (ICIP)*, 168–172.

Chowdhury, A.; Kassem, H.; Padoy, N.; Umeton, R.; and Karargyris, A. 2022. A Review of Medical Federated Learning: Applications inOncology andCancer Research. 3–24.

Dai, Q.; Wei, D.; Liu, H.; Sun, J.; Wang, L.; and Zheng, Y. 2024. Federated Modality-Specific Encoders and Multimodal Anchors for Personalized Brain Tumor Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2): 1445–1453.

Dai, Y.; Chen, Z.; Li, J.; Heinecke, S.; Sun, L.; and Xu, R. 2023. Tackling Data Heterogeneity in Federated Learning with Class Prototypes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6): 7314–7322.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4690–4699.

Feng, C.-M.; Yan, Y.; Wang, S.; Xu, Y.; Shao, L.; and Fu, H. 2023a. Specificity-Preserving Federated Learning for MR Image Reconstruction. *IEEE Transactions on Medical Imaging*, 42(7): 2010–2021.

Feng, T.; Bose, D.; Zhang, T.; Hebbar, R.; Ramakrishna, A.; Gupta, R.; Zhang, M.; Avestimehr, S.; and Narayanan, S. 2023b. FedMultimodal: A Benchmark for Multimodal Federated Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 40354045.

Gu, X.; Guo, Y.; Li, Z.; Qiu, J.; Dou, Q.; Liu, Y.; Lo, B.; and Yang, G.-Z. 2022. Tackling Long-Tailed Category Distribution Under Domain Shifts. In *European Conference on Computer Vision*, 727–743.

Hu, M.; Cao, Y.; Li, A.; Li, Z.; Liu, C.; Li, T.; Chen, M.; and Liu, Y. 2024. FedMut: Generalized Federated Learning via Stochastic Mutation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11): 12528–12537.

Huang, W.; Ye, M.; Shi, Z.; Li, H.; and Du, B. 2023. Rethinking Federated Learning with Domain Shift: A Prototype View. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16312–16322.

Huang, Y.; Chu, L.; Zhou, Z.; Wang, L.; Liu, J.; Pei, J.; and Zhang, Y. 2021. Personalized Cross-Silo Federated Learning on Non-IID Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9): 7865–7873.

Jiang, M.; Wang, Z.; and Dou, Q. 2022. HarmoFL: Harmonizing Local and Global Drifts in Federated Learning on Heterogeneous Medical Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1): 1087–1095.

Khoa, T. A.; Trac, N. D.; Tinh, V. P.; Nam, N. H.; Dang, D. N. M.; Son, H. H.; and Lam, P. D. 2024. Safety Is Our Friend: A Federated Learning Framework Toward Drivers State and Behavior Detection. *IEEE Transactions on Computational Social Systems*, 11(2): 2340–2358.

Li, M.; Cheung, Y.-m.; and Lu, Y. 2022. Long-Tailed Visual Recognition via Gaussian Clouded Logit Adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6929–6938.

Li, Q.; He, B.; and Song, D. 2021. Model-Contrastive Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10713–10722.

Li, X.; Gong, Y.; Huang, K.; and Niu, Z. 2023a. Over-the-Air Integrated Sensing, Communication, and Computation in IoT Networks. *IEEE Wireless Communications*, 30(1): 32–38.

Li, Z.; Shang, X.; He, R.; Lin, T.; and Wu, C. 2023b. No Fear of Classifier Biases: Neural Collapse Inspired Federated Learning with Synthetic and Fixed Classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 5319–5329.

Liu, X.; Hou, S.; Liu, S.; Ding, W.; and Zhang, Y. 2023. Attention-based multimodal glioma segmentation with multi-attention layers for small-intensity dissimilarity. *Journal of King Saud University - Computer and Information Sciences*, 35(4): 183–195.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and Arcas, B. A. y. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1273–1282.

Ouyang, X.; Shuai, X.; Zhou, J.; Shi, I. W.; Xie, Z.; Xing, G.; and Huang, J. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 324337.

Ouyang, X.; Xie, Z.; Fu, H.; Cheng, S.; Pan, L.; Ling, N.; Xing, G.; Zhou, J.; and Huang, J. 2023. Harmony: Heterogeneous Multi-Modal Federated Learning through Disentangled Model Training. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, 530–543.

Poria, S.; Chaturvedi, I.; Cambria, E.; and Hussain, A. 2016. Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 439–448.

Ren, J.; Yu, C.; sheng, s.; Ma, X.; Zhao, H.; Yi, S.; and Li, h. 2020. Balanced Meta-Softmax for Long-Tailed Visual

Recognition. In *Advances in Neural Information Processing Systems*, volume 33, 4175–4186.

Salehi, B.; Gu, J.; Roy, D.; and Chowdhury, K. 2022. FLASH: Federated Learning for Automated Selection of High-band mmWave Sectors. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, 1719–1728.

Shenaj, D.; Fanì, E.; Toldo, M.; Caldarola, D.; Tavera, A.; Michieli, U.; Ciccone, M.; Zanuttigh, P.; and Caputo, B. 2023. Learning Across Domains and Devices: Style-Driven Source-Free Domain Adaptation in Clustered Federated Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 444–454.

Shi, J.; Zheng, S.; Yin, X.; Lu, Y.; Xie, Y.; and Qu, Y. 2024. CLIP-Guided Federated Learning on Heterogeneity and Long-Tailed Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13): 14955–14963.

Tan, J.; Lu, X.; Zhang, G.; Yin, C.; and Li, Q. 2021. Equalization Loss v2: A New Gradient Balance Approach for Long-Tailed Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1685–1694.

Tan, Y.; Long, G.; LIU, L.; Zhou, T.; Lu, Q.; Jiang, J.; and Zhang, C. 2022. FedProto: Federated Prototype Learning across Heterogeneous Clients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8): 8432–8440.

Tao, Y.; Sun, J.; Yang, H.; Chen, L.; Wang, X.; Yang, W.; Du, D.; and Zheng, M. 2023. Local and Global Logit Adjustments for Long-Tailed Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 11783–11792.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5265–5274.

Wang, T.; Zhu, Y.; Chen, Y.; Zhao, C.; Yu, B.; Wang, J.; and Tang, M. 2022. C2AM Loss: Chasing a Better Decision Boundary for Long-Tail Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6980–6989.

Xiao, Z.; Chen, Z.; Liu, L.; Feng, Y.; Wu, J.; Liu, W.; Zhou, J. T.; Yang, H. H.; and Liu, Z. 2024. FedLoGe: Joint Local and Generic Federated Learning under Long-tailed Data. arXiv:2401.08977.

Xueming Yan, S. J., Haiwei Xue; and Liu, Z. 2022. Multimodal Sentiment Analysis Using Multi-tensor Fusion Network with Cross-modal Modeling. *Applied Artificial Intelligence*, 36(1): 2000688.

Ye, R.; Ni, Z.; Xu, C.; Wang, J.; Chen, S.; and Eldar, Y. C. 2023. FedFM: Anchor-Based Feature Matching for Data Heterogeneity in Federated Learning. *IEEE Transactions on Signal Processing*, 71: 4224–4239.

Zeng, Y.; Liu, L.; Liu, L.; Shen, L.; Liu, S.; and Wu, B. 2023. Global Balanced Experts for Federated Long-Tailed Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4815–4825.

Zhang, J.; Liu, Y.; Hua, Y.; and Cao, J. 2024. FedTGP: Trainable Global Prototypes with Adaptive-Margin-Enhanced Contrastive Learning for Data and Model Heterogeneity in Federated Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15): 16768–16776.

Zhang, L.; Lei, X.; Shi, Y.; Huang, H.; and Chen, C. 2023. Federated Learning for IoT Devices With Domain Generalization. *IEEE Internet of Things Journal*, 10(11): 9622–9633.

Zheng, H.; Gao, M.; Chen, Z.; and Feng, X. 2021. A Distributed Hierarchical Deep Computation Model for Federated Learning in Edge Computing. *IEEE Transactions on Industrial Informatics*, 17(12): 7946–7956.

Zuo, H.; Liu, R.; Zhao, J.; Gao, G.; and Li, H. 2023. Exploiting Modality-Invariant Feature for Robust Multimodal Emotion Recognition with Missing Modalities. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.