Incomplete Multimodal Federated Learning via Masking and Contrasting Prototypes

Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, Chaochao Chen, *Senior Member, IEEE*, Liang Hu, and Longbing Cao, *Senior Member, IEEE*

***Abstract*—In real-world scenarios, random modality missingness in multimodal federated learning (mFL) poses a significant challenge, diminishing the performance of global model inference. However, existing mFL methods are predominantly limited to simple scenarios that typically involve participant clients restricted to either a single modality or multimodal clients with complete modalities. They employ modality-specific encoders on each client and train modality fusion modules on the server, leading to severe task drift between clients and server, and struggling to generalize effectively in intricate modality-missing scenarios. To this end, we present a novel mFL framework to alleviate the task drift and performance degradation resulting from modality missingness during both training and inference. Inspired by prototype learning using the highly generalized proxy of specific information, we elaborately construct a prototype library to enhance FedAvg-based federated learning (FL). Naturally, we utilize prototypes as masks representing missing modalities to compensate for the missingness of modality information, formulating a task-calibrated training loss and devising a model-agnostic modality-incomplete inference strategy. In addition, a proximal term based on prototype contrastive learning is constructed to integrate interclient global information into each client, therefore enhancing local training. We conduct extensive experiments to evaluate our mFL framework, demonstrating its state-of-the-art performance across a series of missingness settings. Specifically, compared with existing mFL methods, our mFL framework improves inference performance under different modality missingness rates during training and by 23.8% during modality-incomplete inference.**

***Index Terms*—Federated learning (FL), modality missingness, multimodal learning, representation prototypes.**

I. Introduction

MULTIMODAL pretrained models have exhibited superior performance in various downstream tasks [1], [2], [3], [4], [5], with the availability of large-scale data being a major contributing factor. However, collecting large-scale multimodal data in practical applications may result in great difficulties, such as privacy leakage, data quality, and data availability. An alternative approach is to adopt a decentralized machine learning paradigm, such as federated learning (FL). In FL [6], [7], [8], distributed clients collaborate to train a global model without sharing their private data. While FL can provide data- and communication-efficient training for multimodal models, it faces practical challenges when dealing with complex real-world multimodal data, especially with the widely existing issue of modality missingness. The presence of missing modalities on clients poses constraints of incomplete data on local training, thereby leading to a significant drop in global model inference performance. Furthermore, the non-independent and non-identically distributed (non-IID) nature of multimodal data among clients [9], [10], [11], [12], [13], [14] sharpens the challenges of modality missingness, highlighting the urgent need to address the issue in practical multimodal data.

Some pioneer works [15], [16], [17], [18] attempt to carry out multimodal FL (mFL) with missing modalities. These approaches train modality-specific encoders for each modality, such as visual encoders for images and language encoders for text, using unimodal or self-supervised training tasks. The local encoders are then either aggregated into a global encoder [16] or a global encoder is trained by aligning the local encoders through knowledge distillation (KD) on public data [15]. However, these approaches necessitate additional training data on the server to train a modality fusion module and downstream task heads to cater to various downstream tasks. Also, they only focus on simple scenarios of modality missingness consisting of unimodal clients and modality-complete multimodal clients, as shown in Fig. 1(a). Notably, it is frequently encountered that practical clients possess incomplete data with missing modalities [18] in more intricate real-world scenarios. A typical illustration is observed in social media platforms where users usually generate three types of data: images only, text only, and image–text pairs, resulting in random modality missingness in multimodal data.

When facing generalization or adaptation to such intricate scenarios, the existing FL frameworks manifest the following main issues.

1) Clients cannot obtain fused representations because only modality-specific encoders are deployed.
2) Severe task drift occurs between clients and server, e.g., image/text classification on clients but visual question answering (VQA) on the server, resulting in inconsistent optimization.

Received 24 October 2024; revised 15 June 2025, 19 October 2025, and 18 January 2026; accepted 19 January 2026. Date of publication 23 March 2026; date of current version 2 September 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62576251 and Grant 62376198 and in part by the National Key Research and Development Program of China under Grant 2022YFB3104700. *(Corresponding author: Duoqian Miao.)*

Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, and Liang Hu are with the Department of Computer Science and Technology, Tongji University, Shanghai 201804, China (e-mail: baogy@tongji.edu.cn; zhangqi_cs@tongji.edu.cn; dqmiao@tongji.edu.cn; gongzx@tongji.edu.cn; lianghu@tongji.edu.cn).

Chaochao Chen is with the School of Computer Science, Zhejiang University, Hangzhou, Zhejiang 310058, China (e-mail: zjuccc@zju.edu.cn).

Longbing Cao is with the School of Computing, Macquarie University, Sydney, NSW 2109, Australia (e-mail: longbing.cao@mq.edu.au).

This article has supplementary downloadable material available at https://doi.org/10.1109/TFUZZ.2019.2933787, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2026.3658522

2162-237X © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

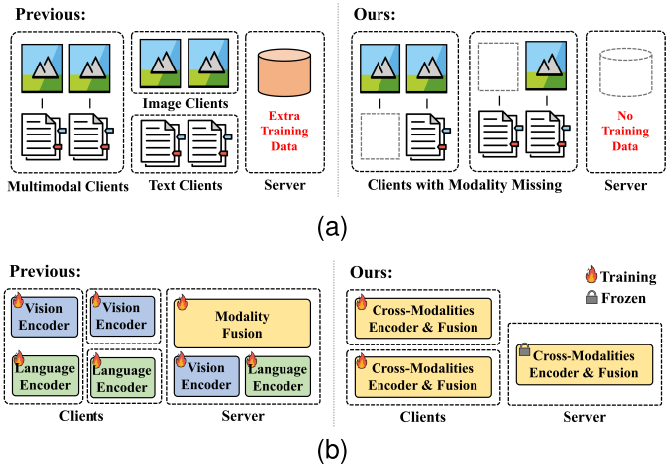


Fig. 1. Illustration of differences between our work and previous works. (a) Difference in task scenarios. Previous works consider two types of clients: unimodal clients and modality-complete multimodal clients. In contrast, our task scenario involves the possibility of partial modalities being missing on each client. In addition, the difference lies in whether the server includes an extra downstream task training set. (b) Differences in model architectures. As shown in (b), the adapted model architectures within the FL framework vary. The previous FL frameworks are restricted to training modality-specific encoders for each modality and then training the modality fusion module on the server. In contrast, ours can develop and train the fusion module on each client. Therefore, it is possible to develop local models with the interactive-encoder architecture encompassing modality encoding and fusion.

3) Lack of specific strategies for addressing modality missingness.
4) The availability of downstream task datasets on servers has also been controversial in FL. These open issues motivate us to explore a new mFL framework targeting real-world modality missingness and challenging fused representations learning.

In this article, we aim to empower the FL framework to handle intricate modality-incomplete scenarios during training and inference. Moreover, we develop modality-interactive encoders and fusion modules for each client, enabling fused representation learning within each client without depending on downstream task datasets on servers [Fig. 1(b)]. To this end, we need to address three primary challenges: CH1—how to handle intricate modality missingness in an interactive encoder?; CH2—how to perform fusion when data is non-IID and modalities are missing?; and CH3—How to ensure inference performance when there are missing modalities?

Specifically, we innovatively utilize the flexible Multiway Transformers [2], [19], [20] to construct the local encoders. By selecting different modality expert networks, multiway transformers can serve as both modality-specific encoders and interactive encoders, thereby structurally adapting to intricate situations of modality missingness (CH1). Inspired by prototype learning in FL [21], [22], [23], [24], [25], [26], we use the representation prototypes, the highly generalized proxy of specific information, to compensate for the missingness of modality information, and integrate global prior information into local training. Specifically, we propose a novel prototype-based mFL framework, termed **P**rototype **M**ask and **C**ontrast for **M**ultimodal **FL** (**PmcmFL**) to achieve fused representation learning with non-IID data and partially missing modalities (CH2). In PmcmFL, we construct and maintain a prototype library in which the prototypes act as masks of the missing modality to compensate for cross-modal fusion and calibrate task drift during training. We also introduce a proximal item based on prototypes to reduce heterogeneity among clients. Subsequently, we introduce elaborate matching algorithms to search for optimal prototypes for inference, which capably tackle missing modalities during inference (CH3).

The main contributions are summarized as follows.

1) Our work stands out as the first attempt to empower our FL framework with the capability to alleviate the global model performance degradation resulting from modality missing during both training and inference.
2) We propose PmcmFL to achieve task-calibrated training and higher inference performance when dealing with modality-incomplete multimodal data.
3) We are the first to adopt representation prototypes to address the intricate situations of modality missingness. Specifically, this is achieved through two operations: prototype masking and prototype contrasting.

We perform comprehensive experiments to evaluate PmcmFL's ability to handle various modality missingness. Extensive results demonstrate that our PmcmFL achieves state-of-the-art performance. Notably, it brings 0.2%–3.7% performance improvements across different modality missingness rates. Regarding inference with missing modalities, PmcmFL also achieves a remarkable 23.8% accuracy improvement. The results confirm the effectiveness and robustness of the proposed PmcmFL in tackling modality missingness.

II. Related Work

A. Multimodal Learning

Multimodal learning has attracted increasing attention from the research community. Multimodal models can be divided into two categories based on whether interaction takes place during modality encoding. The model with dual-encoder architecture [27] uses separate encoders for each modality, with an extra fusion module to conduct modality interaction. On the contrary, the models with interactive-encoder architecture [2], [28], [29] process different modalities as input and conduct modality interactions via fusing tokens. Previous multimodal models in FL [15] have used a dual-encoder architecture, which is not suitable for tasks that require modality fusion.

B. Modality Missingness

Modality missingness is an unavoidable problem in real multimodal tasks, and it is a great challenge to the robustness of multimodal models. To overcome the modality missingness issue in multimodal learning, many methods have been developed. SMIL [30] trains a feature reconstruction network using a meta-learning algorithm. MMIN [31] learns robust multimodal representations by training cascade residual autoencoders. ShaSpec [32] generates features of missing modalities using a shared encoder. Gradient-guided modality decoupling [33] is designed to decouple the dependency on dominating modalities. In addition, some studies have explored the problem of dealing with modal absence in specific areas, such as computer vision [34], [35], medical image [36], recommender systems [37], and multimodal learning [38],

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

[39]. However, these methods introduce additional branching networks and complex training algorithms for the models, which cannot be directly scaled to FL [6], [40].

C. FL With Prototype

Prototype refers to the centroid of the instances belonging to the identical class [41], which has been widely used to solve various challenging issues in FL [22], [42], [43], [44], [45], [46], [47]. Prototypes can act as client–server communication protocol [48], solve domain shift issue [23], [49], [50], handle heterogeneous data [24], [51], [52], and benefit distributed representation learning [53]. Some other FL frameworks [25], [26] without mentioning prototypes still leverage the idea of prototypes to address corresponding issues. Our work primarily utilizes prototypes to address modality missingness issues.

D. Multimodal FL

The mFL [54], [55], [56], [57], [58], [59], [60] is a novel research problem with broad applications in medicine [57], [61], [62], multimodal large language models [63], [64], recommendation systems [65], and other fields [66]. The key challenge lies in how to perform modality interactions when there are missing modalities for each client. FedIoT [16] trains autoencoders for each modality on every client and gives more aggregation weight to multimodal clients. FedMSplit [17] constructs a dynamic graph for adaptively selecting multimodal client models where some modalities may be missing. CreamFL [15] conducts cross-modal interaction using intermodal and intramodal contrast. FedMM [62] employs modality dropout to promote weight alignment processes. MFCPL [67] leverages cross-modal regularization, contrastive mechanisms, and feature alignment (FA) to mitigate severely missing modalities. However, these studies only focus on scenarios with either unimodal clients or multimodal clients with complete modality availability. In contrast, clients where partial data modalities are missing are far more prevalent in real-world settings. FedMultimodal [18] is the first attempt to address this issue by zero-masking the partially missing modalities. We consider scenarios similar to FedMultimodal but with a focus on an efficient and versatile method to address modality missingness.

III. Methodology

A. Overview

We consider an mFL setting including N multimodal clients. Without loss of generality, we take images and text as instances of multimodal data, and it readily extends to other modalities. We discuss extensions of our PmcmFL to other tasks in the Supplementary Materials. We decompose the global model into three functional components: a representation encoder $\mathcal{E}$, a deep fusion layer $\mathcal{F}$, and a task head $\mathcal{G}$. In each image–text pair, the image or text will be missing with a probability of ρ. In this case, all the images can be denoted as $\mathcal{I}_n = \{(i_{nk}, y_{nk})\}_{k=1}^{|\mathcal{I}_n|}$, where i_{nk} is the kth image on the nth client and y_{nk} is its corresponding label. Similarly, all the text can be denoted as $\mathcal{T}_n = \{(t_{nk}, y_{nk})\}_{k=1}^{|\mathcal{T}_n|}$, and all the image–text pairs can be denoted as $\mathcal{M}_n = \{(i_{nk}, t_{nk}, y_{nk})\}_{k=1}^{|\mathcal{M}_n|}$. Furthermore, each client develops a local model $(\mathcal{E}_n, \mathcal{F}_n, \mathcal{G}_n), n \in [1, \ldots, N]$, with the same architecture as the global model.

As shown in Fig. 2, PmcmFL first constructs or updates the prototype library on the server (operation ①) at the beginning of each federated communication round. Subsequently, the prototype library and the global model are broadcast to all participants. Then, each participating client performs local training and then updates local prototypes. Using prototypes as masks of missing modalities, a task-calibrated training loss will supervise the local training (operation ②). Additionally, to tackle the non-IID issue and the resulting client drift, a prototype-based contrastive loss is used (operation ④). Finally, all participating clients transmit their local models and local prototypes back to the server. Then, the global model is updated by aggregating those local models. During inference, the prototype library also assists inference with missing modalities by transmitting prototypes to the global model (operation ③).

For local training, we employ multiway transformer as encoder $\mathcal{E}$. This encoder [2], [19], [20] handles intricate situations of modality missingness by switching among various encoding modes. When only image or text is input, the encoder functions as a modality-specific encoder, like ViT [68] or BERT [69]. When inputting an image–text pair, the encoder functions as an interactive encoder, enabling the representation of one modality to fuse information from the other modality by using cross-modal attention. This ability to switch back and forth based on input modalities enables the encoder to handle intricate missingness. We provide more discussion of the multiway transformer in the Supplementary Materials.

We provide the detailed explanations for each part of our PmcmFL in Sections III-B–III-D and outline it as Algorithm 1.

B. Constructing Prototype Library

Prototypes compact the semantics of data within the same class. Considering communication overhead in FL, low-dimensional representations are used to construct the prototype library. As for the compacting strategy, we naively use the class centroids as the prototypes.

1) Selecting Hidden Representations: For $(i_{nk}, t_{nk}) \in \mathcal{M}_n$, we denote $\mathcal{H}_{nk} = \mathcal{E}_n(i_{nk}, t_{nk}) \in \mathcal{R}^{m \times d}$ as the output of the multiway transformer encoder with the number of output tokens m and the dimension of each token d. The image CLS token $h_{nk}^{(i)}$ and text CLS token $h_{nk}^{(t)}$ are extracted from $\mathcal{H}_{nk}$. We select the two CLS tokens along with the fused representation, i.e., $h_{nk}^{(f)} = \mathcal{F}_n(h_{nk}^{(i)}, h_{nk}^{(t)})$, for constructing prototypes.

2) Construction and Update: In the prototype library, we construct and maintain three types of prototypes: image prototypes $\mathcal{P}_{\mathcal{I}}$, text prototypes $\mathcal{P}_{\mathcal{T}}$, and fusion prototypes $\mathcal{P}_{\mathcal{F}}$. All prototypes are initialized with zero tensors or noise tensors. Subsequently, after local training in each federated communication round, all participating clients construct local prototypes by computing class centroids. Finally, the global prototypes are aggregated from the local prototypes with numbers of client

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

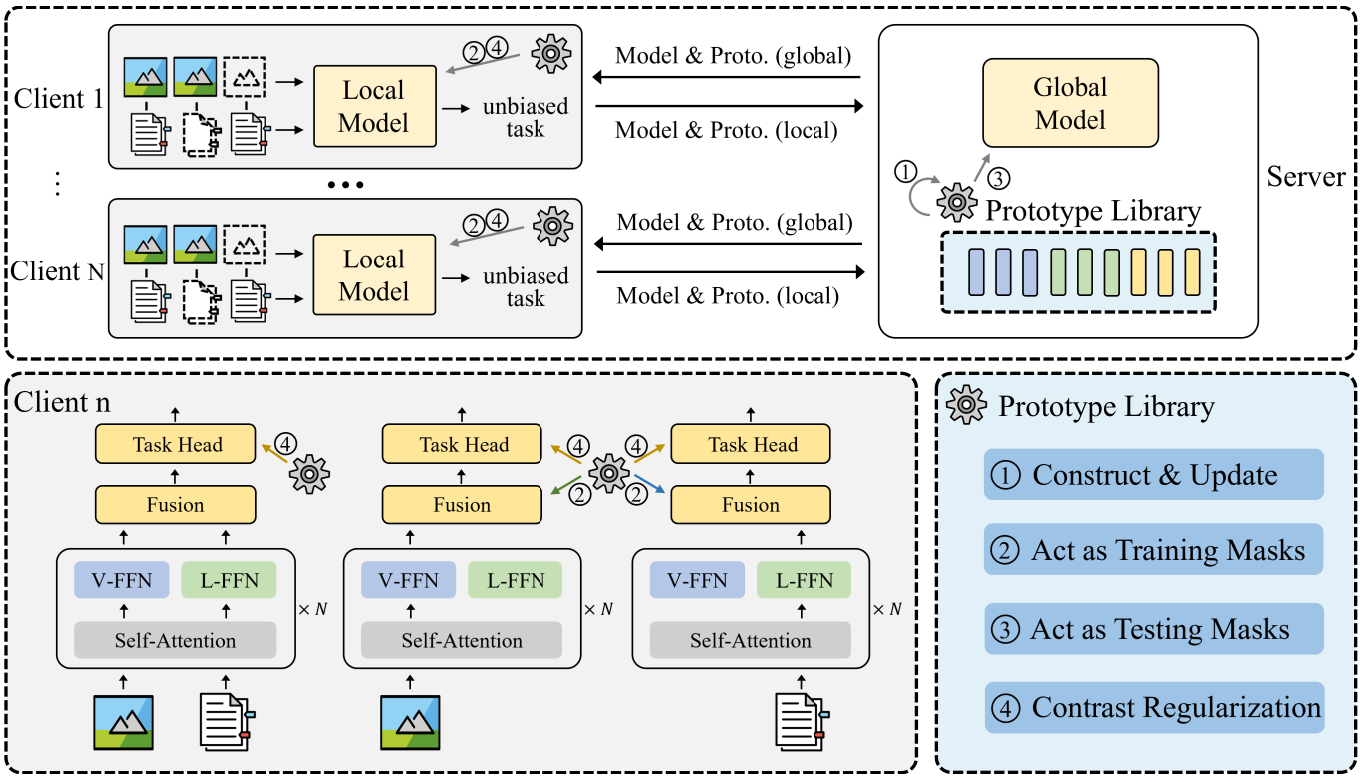


Fig. 2. Overview of the proposed PmcmFL. We first choose low-dimensional representations to construct the prototype library (in Section III-B). Based on this, prototypes act as masks to aid training and inference with missing modalities (in Section III-C). Furthermore, we utilize prototypes to construct contrastive loss for better learning of fused representations (in Section III-D).

samples as weights. The construction process is formalized

$$\mathcal{P}_{\mathcal{I}_n^j} = \sum_{k \in C_n^j} \frac{h_{nk}^{(i)}}{\left|C_n^j\right|}, \quad \mathcal{P}_{\mathcal{I}}{}^j = \sum_{n \in P_c} \frac{|\mathcal{I}_n|}{|\mathcal{D}|} \mathcal{P}_{\mathcal{I}_n^j} \tag{1}$$

$$\mathcal{P}_{\mathcal{T}_n^j} = \sum_{k \in C_n^j} \frac{h_{nk}^{(t)}}{\left|C_n^j\right|}, \quad \mathcal{P}_{\mathcal{T}}{}^j = \sum_{n \in P_c} \frac{|\mathcal{T}_n|}{|\mathcal{D}|} \mathcal{P}_{\mathcal{T}_n^j} \tag{2}$$

$$\mathcal{P}_{\mathcal{F}_n^j} = \sum_{k \in C_n^j} \frac{h_{nk}^{(f)}}{\left|C_n^j\right|}, \quad \mathcal{P}_{\mathcal{F}}{}^j = \sum_{n \in P_c} \frac{|\mathcal{M}_n|}{|\mathcal{D}|} \mathcal{P}_{\mathcal{F}_n^j} \tag{3}$$

where C_n^j denotes the data of class j on the nth client, P_c denotes the set of participating clients, and $\mathcal{D}$ denotes all data of participating clients. For $* \in [\mathcal{I}, \mathcal{T}, \mathcal{F}]$, $\mathcal{P}_{*_n^j}$ denotes the local prototypes and $P_*{}^j$ denotes the global prototypes. Please note that the computation of the prototypes can be optimized by removing outliers. Specifically, we can calculate the Z-score for all data points within the same class and remove certain outliers either by setting a threshold or by eliminating a fixed proportion of the highest scoring values. However, we still adopt the simple definition above, mainly demonstrating the effectiveness of our framework.

C. Prototypes as Masks of Missing Modalities

When missing modalities occur, previous work [18] masks missing data with zero tensors. In PmcmFL, prototypes are employed as masks for missing modality representations, thereby compensating for missing information. Based on this, a task-calibrated training loss and a model-agnostic unimodal inference strategy are proposed.

1) Task-Calibrated Training Loss: As shown in Fig. 2, during local training with complete image–text pairs, the training loss is defined as

$$L_{task} = \mathcal{L}\left(\mathcal{G}_n\left(\mathcal{F}_n\left(h_{nk}^{(i)}, h_{nk}^{(t)}\right)\right), y_{nk}\right) \tag{4}$$

where $\mathcal{L}$ denotes the task loss function. For classification tasks, cross entropy is commonly used. In the case of local training with text missing, a task-calibrated training loss is constructed using the text prototypes as masks

$$L_{task} = \mathcal{L}\left(\mathcal{G}_n\left(\mathcal{F}_n\left(h_{nk}^{(i)}, \mathcal{P}_{\mathcal{T}}{}^j\right)\right), y_{nk}\right) \tag{5}$$

where $\mathcal{P}_{\mathcal{T}}{}^j$ has the same class as y_{nk}. Similarly, the task-calibrated training loss is defined when images are missing

$$L_{task} = \mathcal{L}\left(\mathcal{G}_n\left(\mathcal{F}_n\left(\mathcal{P}_{\mathcal{I}}{}^j, h_{nk}^{(t)}\right)\right), y_{nk}\right). \tag{6}$$

With supervision from the task-calibrated loss, the entire model is trained on the original multimodal task, even when there are missing modalities on clients. On the contrary, using zero masks will result in a task-drifted loss. For example, in the case where text is missing, the training loss with zero masks is denoted as $\mathcal{L}(\mathcal{G}_n(\mathcal{F}_n(h_{nk}^{(i)}, \mathbf{0})), y_{nk})$. Essentially, this is a unimodal training task that depicts the relationship between the unimodal representation $h_{nk}^{(i)}$ and the ground truth y_{nk}, leading to inconsistent optimization for the model training.

2) Model-Agnostic Unimodal Inference: During inference with missing modalities, prototypes are employed as masks to assist the global model. Unlike training, where labels are available, we need to find the corresponding cross-modal prototypes according to embedded representations. To this end, a matching function $m(\cdot)$ is established from hidden representations to

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 Training Algorithm of PmcmFL

Input: Number of federated communication rounds T, number of clients C, number of local epochs E, server model, local model $(\mathcal{E}_n, \mathcal{F}_n, \mathcal{G}_n)$, local learning rate η, dataset $\mathcal{D}_n = (\mathcal{M}_n, \mathcal{I}_n, \mathcal{T}_n)$ of the n-th client and fraction of clients s that are selected to perform computation in each round.

Output: The final server model's parameters ω^T

ServerExecutes:

Initialize ω^0 randomly;
Initialize global prototypes $\mathcal{P}_{\mathcal{I}-local}, \mathcal{P}_{\mathcal{T}-local}, \mathcal{P}_{\mathcal{F}-local}$ with zero tensors;
for $t = 1, 2, \ldots, T$ **do**
  Compute global prototypes $\mathcal{P}_{\mathcal{I}}, \mathcal{P}_{\mathcal{T}}, \mathcal{P}_{\mathcal{F}}$ according to Equation 1 2 3;
  $\mathcal{P}_{\mathcal{I}-local}, \mathcal{P}_{\mathcal{T}-local}, \mathcal{P}_{\mathcal{F}-local} \leftarrow [\,], [\,], [\,]$;
  $S_t \leftarrow$ random set of $\max(s \cdot C, 1)$ clients;
  for *each client n in* S_t **do**
    send the global model's parameters ω^{t-1} and global prototypes $\mathcal{P}_{\mathcal{I}}, \mathcal{P}_{\mathcal{T}}, \mathcal{P}_{\mathcal{F}}$ to client n;
    $\omega_n^t, \mathcal{P}_{\mathcal{I}n}, \mathcal{P}_{\mathcal{T}n}, \mathcal{P}_{\mathcal{F}n} \leftarrow$ **ClientLocalTraining** $(n, t, \omega^{t-1}, \mathcal{P}_{\mathcal{I}}, \mathcal{P}_{\mathcal{T}}, \mathcal{P}_{\mathcal{F}})$;
    $\mathcal{P}_{\mathcal{I}-local}, \mathcal{P}_{\mathcal{T}-local}, \mathcal{P}_{\mathcal{F}-local} \leftarrow \mathcal{P}_{\mathcal{I}-local} + \mathcal{P}_{\mathcal{I}n}, \mathcal{P}_{\mathcal{T}-local} + \mathcal{P}_{\mathcal{T}n}, \mathcal{P}_{\mathcal{F}-local} + \mathcal{P}_{\mathcal{F}n}$;
  end
  Local models are aggregated into the global model: $\omega^t = \sum_{n \in S_t} \frac{|\mathcal{D}_n|}{|\mathcal{D}|} \omega_n^t$
end

ClientLocalTraining $(n, t, \omega^{t-1}, \mathcal{P}_{\mathcal{I}}, \mathcal{P}_{\mathcal{T}}, \mathcal{P}_{\mathcal{F}})$:

for *epoch* $i = 1, 2, \cdots E$ **do**
  Local update $\omega_n^{(t,i)} \leftarrow \omega_n^{(t,i-1)} - \eta \nabla L(\mathcal{D}_n, \mathcal{P}_{\mathcal{I}}, \mathcal{P}_{\mathcal{T}}, \mathcal{P}_{\mathcal{F}}; \omega_n^{(t-1,i-1)})$ according to Equation 9;;
end
Compute local prototypes $\mathcal{P}_{\mathcal{I}n}, \mathcal{P}_{\mathcal{T}n}, \mathcal{P}_{\mathcal{F}n}$ according to Equation 1 2 3;
Return $\omega_n^t, \mathcal{P}_{\mathcal{I}n}, \mathcal{P}_{\mathcal{T}n}, \mathcal{P}_{\mathcal{F}n}$;

prototypes. The learning objective of $m(\cdot)$ is minimizing the distance between representations and prototypes for the same modality and classes. Using ω to parameterize the matching function, the learning objective can be formally expressed as

$$\omega^* = \min_{\omega} \sum_{n,j} Dis\left(m\left(h_{nk}^j|\omega\right), \mathcal{P}^j\right) \tag{7}$$

where h_{nk}^j and $\mathcal{P}^j$ are the matched representations and prototypes and $Dis(\cdot)$ is the distance metric (typically L2 distance). We employ specific matching functions for each modality.

Based on the matching function $m(\cdot)$, we can find the ideal prototype for each representation as follows. Without loss of generality, assume that the global model only uses images for inference. First, the image i_{nk} is input for computing the image representation $h_{nk}^{(i)}$. Sequentially, the most ideal image prototype $\mathcal{P}_{\mathcal{I}}{}^j$ is determined using this matching function $\mathcal{P}_{\mathcal{I}}{}^* = m(h_{nk}^{(i)})$. Following this, the ideal text prototype $\mathcal{P}_{\mathcal{T}*}$ is determined according to the class of $\mathcal{P}_{\mathcal{I}}{}^*$. Finally, the image representation $h_{nk}^{(i)}$ and the text prototype $\mathcal{P}_{\mathcal{T}*}$ are input into the subsequent modules.

To obtain the matching function $m(\cdot)$, we propose both model-free and model-based matching. Model-free matching involves matching the closest prototype for each representation using a distance metric, such as L1, L2 distance, or cosine similarity. Model-based matching involves training tiny models, treating the search for cross-modal prototypes as either a classification or a retrieval task. These tiny models are trained only in the final round of federated communication, incurring negligible overhead to the FL framework. For the classification task, we train a shallow MLP-Classifier using cross-entropy loss. For the retrieval task, inspired by DALLE·2 [70], we sequentially use CLIP loss [27] and mean square error (mse) loss to train an MLP-Prior to achieve transfer and retrieval. We attempt three aggregation strategies for tiny models from different clients: selecting the model with maximum client samples, model aggregation, and model ensemble.

Inspired by MixUp [71], we propose ProtoMix to enrich the semantic information of prototypes for inference. ProtoMix linearly combines the top-k related prototypes with weights derived from applying the softmax function to distance metrics, classification probabilities, or retrieval match scores.

D. Prototypes as Contrastive Representation Targets

Due to the non-IID issue, the learned representation distributions in latent space significantly differ among clients (i.e., client drift), leading to an uncoordinated model aggregation. To achieve model aggregation without conflicts, all clients must theoretically have similar local representation distributions. In PmcmFL, prototypes are used as contrastive representation targets to guide clients in learning fused representation distributions clustered by class. For this reason, we introduce a proximal term based on the unidirectional CLIP contrastive loss to regularize client training. Sampling a batch $\mathcal{B}$ from the local dataset, the proximal term can be denoted as

$$L_{contr} = -\frac{1}{|\mathcal{B}|} \sum_{s=1}^{|\mathcal{B}|} \log \frac{\exp\left(h_s^\top \cdot \mathcal{P}_s/\tau\right)}{\sum_{i=1}^{|\mathcal{B}|} \exp\left(h_s^\top \cdot \mathcal{P}_i/\tau\right)} \tag{8}$$

where h_s is the sth representation in batch, $\mathcal{P}_s$ is its corresponding prototype, and τ is a temperature factor. The prototypes gradually converge as the global model also converges. This proximal term works by pulling representations closer to the prototypes of the same class and pushing representations away from prototypes of different classes.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

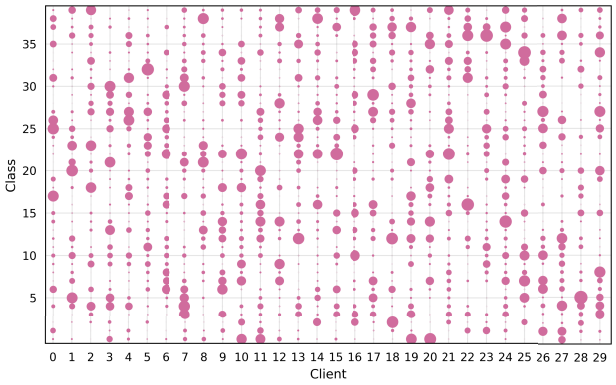


Fig. 3. Non-IID data distribution among clients. We select 40 categories from all 30 clients for visualization, with the size of the dots representing the amount of data for each category.

Thus, representations are around the corresponding global prototypes.

Since prototypes are the centroids of class representations, maintaining a certain distance (i.e., fine-grained semantic difference) from each representation, we use CLIP loss to constrain their similarities rather than relying on mse to constrain their distances. Although the prototype-based proximal term can theoretically be constructed on image representations, text representations, and fused representations separately, we only utilize the last one due to the difficulty in coordinating multiple proximal terms. In this case, the total training loss for each client is the sum of task loss and contrastive loss of fused representations

$$L = L_{task} + \gamma L^{f}_{contr} \tag{9}$$

where γ is the weighting hyperparameter for L^{f}_{contr}.

IV. Experimental Setting

A. Dataset

Following CreamFL [15], we use the challenging VQAv2 dataset [72] to evaluate our PmcmFL. VQAv2 is an authoritative multimodal classification dataset, with images sourced from MSCOCO [73], meticulously annotated with text and labels for each image. To enable better and more efficient evaluation, we performed a simple filtering of the dataset. Specifically, we filtered out data in the training set where the labels occurred infrequently. This process resulted in a dataset with 310 categories, comprising 64 000 training samples and 5000 testing samples, which we call tinyVQA. We distribute the training data to 30 clients, employing the Dirichlet distribution ($\alpha = 0.1$) for non-IID data partition [74]. This setting causes severe non-IID issues between clients, leading to client drift, as visualized in Fig. 3. We also considered different heterogeneity metrics and numbers of clients in the robustness evaluation. The above setting is a realistic simulation of data distribution in real-world scenarios, and we provide more description of the dataset in the Supplementary Materials.

B. Baselines

The proposed PmcmFL aims to address modality missingness and non-IID issues in real-world scenarios. Therefore, the federated frameworks selected as baselines for evaluation should have targeted strategies to handle these two issues. Some of the methods we mentioned earlier, such as FedIoT and CreamFL, incorporate both targeted strategies. However, these two federated frameworks are only suitable for simple missingness scenarios, where there are only unimodal clients and modality-complete multimodal clients. Thus, the comparison with them can only be evaluated in such simple scenarios.

To evaluate intricate missingness scenarios, we combine existing methods that address modality missingness in FL with methods that tackle non-IID issues as baselines for evaluation. On the one hand, to handle the modality missingness in FL training, data cleaning is the most direct approach, which involves discarding incomplete data and only using complete data to train local models. We refer to this method as **Ignore Missingness**. FedMultimodal [18] involves a concise and effective method for handling modality missingness in FL training. It uses zero tensors to mask missing data or modalities, thereby leveraging the missing data. We refer to this method as **Zero Mask**. An approximate solution to Zero Mask is the **Noise Mask**, which uses noise tensors to mask missing modalities. We further propose a variant of our PmcmFL as a baseline, which is named PmmFL (PmcmFL without contrast) and uses **Prototype Mask** to handle modality missingness. On the other hand, non-IID issues are widely studied in FL research. We select FedProx [75], FedIoT [16], FedPAC [25], and FedHKD [26] as targeted strategies for non-IID, all of which are FL frameworks based on FedAvg [6]. FedProx introduces a proximal term, which restricts the range of updates to local model weights, ensuring that local models do not deviate significantly from the global model. FedIoT considers the non-IID modalities issue by providing additional aggregation weights to multimodal clients, thus alleviating client drift. Here, we make a slight modification. We assign aggregation weights based on the amount of complete multimodal data, i.e., the weight for client n is proportional to $|\mathcal{M}_n|$. FedPAC proposes FA to force the representation spaces of clients and the server closer, and FedHKD uses hyperknowledge distillation (HKD) to align clients and the server. Both FedPAC and FedHKD are prototype-based FL frameworks. We provide more details of baselines in the Supplementary Materials.

C. Implementation Details

Regarding PmcmFL, we utilize the same architecture as BEIT3-base [2] along with its pretrain weight for the encoder $\mathcal{E}(\cdot)$. Due to pretraining data not including VQAv2, the use of pretraining weight is only for fast convergence and does not affect the reliability of the evaluation results. For the deep fusion layer $\mathcal{F}$, we adopt an MLP that takes the image CLS token and text CLS token generated by the encoder $\mathcal{E}$ as input and outputs a fused representation. For the classification task head $\mathcal{G}$, the fused representation is projected to probability distributions using another MLP. The image CLS tokens, text CLS tokens, and fused representations are used to construct the image prototypes, text prototypes, and fused prototypes, respectively. We provide more details in the Supplementary Materials.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

TABLE I

COMPARISON OF PMCMFL WITH BASELINES UNDER VARIOUS TRAINING MODALITY MISSINGNESS. MODALITY MISSINGNESS OCCURS ONLY IN THE TRAINING PHASE, AND THE TESTING DATA IN THE INFERENCE PHASE ARE MODALITY-COMPLETE. NOTE THAT FEDAVG+ZERO MASK IS FEDMULTIMODAL'S STRATEGY FOR MODALITY MISSINGNESS. THE RESULTS ARE AN AVERAGE VALUE OF FIVE RUNS. * DENOTES THAT OUR PMCMFL'S PERFORMANCE IS SIGNIFICANTLY BETTER THAN EXISTING BASELINES (PAIRED T-TEST, $p < 0.05$)

Frameworks	Missingness Handling	Accuracy (%) on 5K Testing Samples ($\rho_{test} = 0$)					Acc@sum
		$\rho_{train} = 0.1$	$\rho_{train} = 0.2$	$\rho_{train} = 0.3$	$\rho_{train} = 0.4$	$\rho_{train} = 0.5$	
FedAvg	Ignore Missingness	56.444	53.936	49.990	45.024	38.898	244.292
FedAvg	Zero Mask	55.360	52.804	49.246	46.378	41.034	244.822
FedAvg	Noise Mask	55.624	52.542	48.164	47.288	41.346	244.964
FedProx	Noise Mask	52.506	49.546	46.502	46.862	41.140	236.556
FedIoT	Noise Mask	55.010	52.180	47.558	47.582	41.838	244.168
FedPAC	Noise Mask	55.028	50.742	46.916	45.930	40.242	238.858
FedHKD	Noise Mask	56.286	53.546	50.544	47.832	43.376	251.584
FedPAC	Prototype Mask (**Ours**)	56.144	54.366	50.556	48.796	46.168	256.030
FedHKD	Prototype Mask (**Ours**)	56.458	53.926	51.568	49.472	46.394	257.818
PmcmFL (**Ours**)	Prototype Mask (**Ours**)	**56.636***	**55.478***	**52.256***	**49.548***	**47.042***	**260.960***

For PmcmFL and all baselines, we set the number of communication rounds to 30, the number of local training rounds to 5, and the learning rate to $3e^{-4}$ using the AdamW optimizer for local training. The hyperparameter γ is searched within $[5.0, 1.0, 0.5, 0.1, 0.01]$ to find the optimal value. For the hyperparameters in baselines, we use the optimal values reported in the corresponding papers. We provide more implementation details in the Supplementary Materials.

D. Evaluation Metric

We use VQAv2-specific accuracy [72], which is robust to interhuman variability, as the accuracy metric. We utilize optimal performance in the last ten communication rounds as the final performance. All results are average values under five runs with different random seeds.

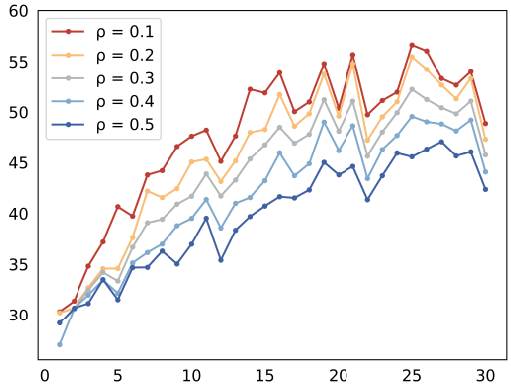


Fig. 4. Performance of the PmcmFL's global model at different federated rounds. The global model gradually converges.

V. EXPERIMENTS AND ANALYSIS

A. Evaluation Under Intricate Modality Missingness

We first evaluate the intricate modality missingness. In this scenario, each client will miss some of the images or text used for local training, with varying missingness rates among different clients. Due to the complexity of this assumption, which makes comprehensive evaluation difficult, we follow the settings of previous research [18] and set the same training modality missingness rate ρ_{train} for all clients. Specifically, the modality missingness rates during the federated training phase ρ_{train} are set to [0.1, 0.2, 0.3, 0.4, 0.5], which covers a range from minor to severe missingness. We first use the FedAvg to pre-evaluate the performance of different missingness handling strategies, i.e., evaluating Ignore Missingness, Zero Mask, and Noise Mask. The optimal strategy obtained from this pre-evaluation will then be applied to other FL frameworks for the subsequent evaluation. We temporarily refrain from applying missingness to the test set and use a modality-complete test set for evaluation. Table I displays the performance of all baselines and our PmcmFL. It can be observed that our PmcmFL achieves noticeable performance improvement over all baselines under training modality missingness rates ρ_{train} ranging from 0.1 to 0.5. We further demonstrate the performance of the PmcmFL's global model at different rounds in Fig. 4. It can be seen that as federated training progresses, the performance of the global model continuously improves and tends to converge.

Comparing three existing modality missingness handling strategies under the FedAvg framework, we observe a similar Acc@sum performance. When the modality missingness rate is low, ignoring incomplete data results in better global model accuracy. This is reasonable because discarding a small portion of data hardly affects the generalization performance of the global model, while pairing images or texts with zeros/noise introduces additional interference to the knowledge already learned by the model, thereby reducing accuracy. However, as the modality missingness rate gradually increases, we find that using Zero Mask or Noise Mask yields better performance. Specifically, a 2.448% performance boost is achieved with 50% modality missingness. This is because, at a higher missingness rate, discarding a large amount of incomplete data would diminish the generalization performance of the global model. At this point, involving the remaining modalities in training allows the model to learn new knowledge from a large number of previously unseen data patterns, thus enhancing its generalization ability. Additionally, since the Noise Mask achieves the best Acc@sum performance in pre-evaluation, we consider it as the best missingness handling baseline and combine it with other FL baselines for subsequent evaluation.

Compared to FedAvg, both FedProx and FedIoT lead to a performance decline. This indicates that imposing constraints on model parameters or improving the aggregation strategy fails to achieve the original effect, further highlighting the uniqueness of the non-IID problem in scenarios with modality

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

TABLE II

COMPARISON OF PMCMFL WITH BASELINES UNDER VARIOUS MODALITY MISSINGNESS. BOTH TRAINING AND INFERENCE OCCUR MODALITY MISSINGNESS WITH THE SAME RATE ρ. THE RESULTS ARE AN AVERAGE VALUE OF FIVE RUNS

Frameworks	Missingness Handling	Accuracy (%) on 5K Testing Samples ($\rho_{train} = \rho_{test} = \rho$)					Acc@sum
		$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	
FedAvg	Ignore Missingness	52.016	47.318	39.250	34.218	26.926	199.728
FedAvg	Zero Mask	52.822	47.624	43.174	38.824	34.658	217.102
FedAvg	Noise Mask	52.268	47.602	41.614	39.862	34.082	215.428
FedProx	Noise Mask	50.382	44.978	40.808	40.240	34.138	210.546
FedIoT	Noise Mask	52.714	47.112	41.266	39.968	34.430	215.490
FedPAC	Noise Mask	52.044	46.094	41.084	38.258	33.872	211.352
FedHKD	Noise Mask	53.774	49.654	43.186	40.962	37.394	224.970
PmcmFL (**Ours**)	Prototype Mask (**Ours**)	**53.810**	**50.606**	**44.916**	**41.014**	**38.264**	**228.610**

missingness. Similarly, FedPAC also leads to a performance decline (a 6.106% drop on Acc@sum). We argue that this is because the FA employed by FedPAC, a strong regularization that narrows the distance between representations and prototypes through mse, disrupts the semantic distinctiveness of representations. On the contrary, FedHKD has a 6.62% performance improvement compared to FedAvg. This is because HKD is a soft alignment that conducts alignment between prototypes rather than aligning data representations with prototypes. Compared to all the baselines, our PmcmFL achieves superior performance (9.376%–24.404% boost on Acc@sum), demonstrating the effectiveness of our Prototype Mask and Contrast. Furthermore, as the modality missingness rate increases, the performance improvement brought about by our PmcmFL tends to be more significant (a 0.192% boost in 10% missing but a 3.666% boost in 50% missing). Compared to PmcmFL's variants, slightly superior performance (3.14-4.93% boost on Acc@sum) indicates that our Prototype Contrast is more generalizable than FA and HKD.

B. Evaluation With Modality-Incomplete Test Set

In the evaluations described above, modality missingness occurs only in the training phase, while the modality is complete in the inference phase. This can be considered an out-of-distribution setting due to the inconsistency between the training and test sets in terms of modality composition. However, in real scenarios, modality missingness also occurs in the inference phase, and it is natural to want to explore how the FL framework performs under different inference missingness rates. To this end, we utilize a modality-incomplete test set to evaluate global models trained in Section V-A. Specifically, we set the modality missingness rates of inference equal to those of training. We further consider a more complex setup in the Supplementary Materials. Regarding the ideal prototype matching strategy, we decide to train classifiers and integrate them using parameter aggregation. Section V-E will give reasons for this decision. Table II displays the evaluation results.

It can still be seen that our PmcmFL achieves optimal performance at all missingness rates. When comparing other methods to Ignoring Missingness, a substantial performance gain is observed at the 0.5 modality missingness level (7.156%–11.338% boost). This demonstrates the sensibility of handling missingness using the Mask. We also observe similar conclusions as in Section V-A.

TABLE III

EVALUATION ON THE SIMPLE SCENARIO. ALL RESULTS OF BASELINES ARE CITED FROM CREAMFL [15]

Methods	Accuracy
FedAvg	52.54
FedIoT	53.06
FedMD	57.43
FedET	59.90
FedGEMS	60.23
reamFL+Avg	58.64
reamFL+IoT	59.64
CreamFL	62.12
PmcmFL(**Ours**)	**65.53**

C. Evaluation Under Simple Modality Missingness

We similarly evaluate our PmcmFL in previously extensively studied missingness scenarios, i.e., unimodal clients and modality-complete multimodal clients, which is a special case of the scenarios we consider. We need to emphasize that this simple scenario is a special case of the intricate scenario we considered above. When the missingness rate of a specific modality on the partial client reaches 100%, our task scenario degenerates into this special case. We follow CreamFL for client setup and take 5000 random samples from the validation set of VQAv2 for testing. The training data and the remaining validation set data are distributed to 35 clients, including 15 multimodal clients, 10 image clients, and 10 text clients. The evaluation is conducted over 3000 of the most frequent answers, and the 5000 testing samples are modality-complete. Table III displays the results.

The results show that our PmcmFL also achieves state-of-the-art performance (a 3.38% boost compared with CreamFL). This suggests that PmcmFL is equally applicable to simple missingness scenarios considered by previous works and outperforms existing methods specifically designed for such scenarios. We believe that the performance improvement mainly stems from the advantage of our PmcmFL, which could train a modality fusion module on the client.

D. Ablation Studies

We investigated the effect of each component of PmcmFL in this section. With the training missingness rate set to $[0.1, 0.2, 0.3, 0.4, 0.5]$, we ablate the Prototype Contrast and Prototype Mask components for federation training. During evaluation, the test set maintains modality integrity without missingness. When removing the Prototype Mask, we use the Noise Mask instead. When ablating these two components simultaneously, PmcmFL degrades to FedAvg combined with the Noise Mask. Table IV displays the ablation results.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
RESULTS OF ABLATION STUDIES UNDER VARIOUS TRAINING MISSINGNESS RATES

Prototype Contrast	Prototype Mask	Accuracy (%) on 5K Testing Samples ($\rho_{test} = 0$)					Acc@sum
		$\rho_{train} = 10\%$	$\rho_{train} = 20\%$	$\rho_{train} = 30\%$	$\rho_{train} = 40\%$	$\rho_{train} = 50\%$	
		55.624	52.542	48.164	47.288	41.346	244.964
✓		55.948	53.588	50.982	47.422	44.454	252.394
	✓	55.326	54.486	50.858	49.290	45.848	255.808
✓	✓	**56.636**	**55.478**	**52.256**	**49.548**	**47.042**	**260.960**

TABLE V
INFERENCE WITH 100% TEXT MODALITY MISSING. ZERO MASK AND NOISE MASK ARE BASELINE METHODS FOR HANDLING MODALITY MISSINGNESS DURING INFERENCE. OTHERS ARE OUR PROTOTYPE MASK WITH VARIOUS IDEAL PROTOTYPE MATCHING ALGORITHMS

Methods	Accuracy (%) on Unimodal Data			
	Mix-1	Mix-10	Mix-20	Mix-best
Zero Mask	0.996	0.996	0.996	0.996
Noise Mask	3.662	3.662	3.662	3.662
L1 Metric	6.358	6.170	6.784	6.784
L2 Metric	5.454	5.238	5.782	6.106
COS Metric	5.524	6.062	6.676	6.938
Classifier(max.)	26.956	27.022	27.022	27.022
Classifier(ens.)	27.090	27.090	27.090	27.090
Classifier(avg.)	**27.502**	27.092	27.092	**27.502**
MLP-Prior(max.)	27.052	19.906	8.742	27.052
MLP-Prior(ens.)	27.260	27.280	27.280	27.280
MLP-Prior(avg.)	5.858	5.702	5.996	6.210

The results show that Prototype Contrast and Prototype Mask contribute to performance improvement, with gains of 7.430% and 11.168% on Acc@sum, respectively. Furthermore, we find that the improvement brought by Prototype Mask is greater than that of Prototype Contrast. This suggests that modality missingness can be more damaging to global model performance and that our method mitigates the effect of missingness on training well. Finally, using both Prototype Contrast and Prototype Mask performs better than using either, suggesting that the two components of our PmcmFL work in harmony and together, rather than being mutually exclusive.

E. Evaluation for Various Prototype Matching Algorithms

Various matching algorithms are designed to find the most ideal prototype. In this section, we evaluate their performance. Without loss of generality, we set the modality missingness rate in the training phase to 0.3, on which the federated training is performed. To intuitively compare various matching algorithms, we let the text or image of the test set be completely missing and construct the inference evaluation with unimodality data. Table V shows the performance when text is completely missing. Zero mask and noise mask serve as baselines for handling inference missingness. For our ProtoMix strategy, we varied the hyperparameter top-k across its entire range of possible values. Here, we present the results for three specific settings: top-k set to 1, 10, and 20, denoted as Mix-1, Mix-10, and Mix-20, respectively. Furthermore, we present the result when the hyperparameter top-k is set to the optimal value, denoted as Mix-best.

According to the results, we can see that the performance achieved by Zero Mask and Noise Mask is poor. This is mainly due to the fact that these mask approaches do not carry any helpful semantic information. In contrast, our Prototype Mask can directly achieve better performance even though we use model-free matching algorithms. However, the ideal prototype matching accuracy achieved by the model-free matching strategy is very low, with only 2.46%, 1.04%, and 1.04% for L1 metric, L2 metric, and COS metric, respectively. This result is reasonable because the prototypes are the mean values of representations and they are not close to each other (or the cosine similarity is not high). Moreover, as the previous study [70] has illustrated, contrastive learning tends to produce disjointed representations, i.e., distances that are not close. Therefore, our Prototype Contrast still has difficulties in bridging the distance between representations and prototypes. This highlights the need to utilize model-based matching algorithms. The results demonstrate that using model-based matching can achieve a performance of up to 27.502%, representing a 23.840% improvement compared to the baselines. This indicates that our method significantly restores the inference performance of the global model, even when one modality is completely missing. We attribute this outstanding performance to the ability of the prototype to introduce accurate semantics. Furthermore, we also observe an anomaly where retrieval-based prototype matching leads to poor performance when using parameter aggregation. We believe that this is due to the fact that the parameter aggregation approach can only be applied to classification supervision, not retrieval supervision.

Observing the effect of ProtoMix under different prototype matching algorithms, we find that it leads to a small performance improvement, suggesting that the fused prototype can provide richer semantics than a single prototype. We discuss this more in the Supplementary Materials.

F. Qualitative Studies

We use t-SNE [76] for visualization to qualitatively analyze the effect of the proposed Prototypes Contrast in our PmcmFL.

Fig. 5(a) visualizes the fused representation learned by three clients. In FedAvg, we observe that some data points are clustered by clients, indicating that each client forms its own distinct representation of latent space. When these local models are aggregated into a global model, the individual latent spaces impede the formation of a unified latent space, resulting in poor performance. In contrast, in our PmcmFL approach, the clustering trend of the data points diminishes. Notably, the blue and red points are evenly mixed together. This suggests that Prototype Contrast reduces heterogeneity between clients, leading to a more similar representation of latent spaces across individual clients.

Fig. 5(b) visualizes the fused representation learned by the global model. We randomly selected samples from three categories in the test set. It can be seen that in FedAvg, points from different categories are mixed together, which is detrimental to the subsequent classification task. In contrast, in our PmcmFL, points from different categories cluster around their respective prototypes due to the ability of the prototypes

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

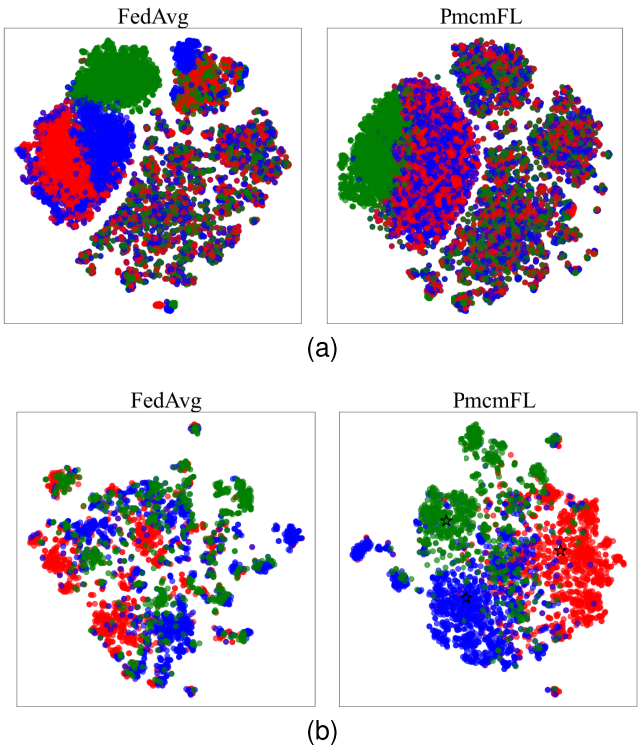


Fig. 5. t-SNE visualization results for qualitative studies between FedAvg and our PmcmFL (a) Representation distribution learned by different clients. Dots of the same color denote representation from the same client. (b) Representation distribution learned by the global model. Dots of the same color denote representations of the same class. The stars represent prototypes for corresponding classes.

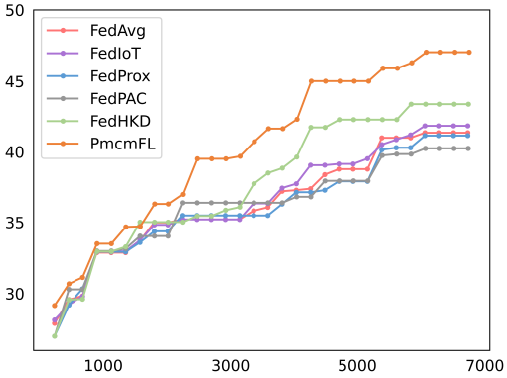


Fig. 6. Communication efficiency. This figure displays the optimal performance of global model relative to the volume of transmitted parameters (M).

to serve as alignment targets to guide the distribution of representations. This validates our starting point of using prototypes to construct contrastive loss.

We provide additional visualization examples in the Supplementary Materials.

G. Overhead and Efficiency of Communication

For FL, the parameter efficiency is an important metric. To evaluate this, communication overhead per federated round and communication efficiency are often considered. The former refers to the number of parameters that need to be transmitted in a single federated communication, while the latter refers to the number of parameters that need to be transmitted to achieve the same performance.

TABLE VI
COMMUNICATION OVERHEAD OF EACH FEDERATED ROUND

Frameworks	Communication Overhead	
	tinyVQA	VQAv2
FedAvg	224.37 MB (+0%)	224.37 MB (+0%)
FedProx	224.37 MB (+0%)	224.37 MB (+0%)
FedIoT	224.37 MB (+0%)	224.37 MB (+0%)
FedPAC	224.62 MB (+0.11%)	226.77 MB (+1.07%)
FedHKD	224.71 MB (+0.15%)	236.55 MB (+5.43%)
PmcmFL	225.09 MB (+0.32%)	231.57 MB (+3.21%)

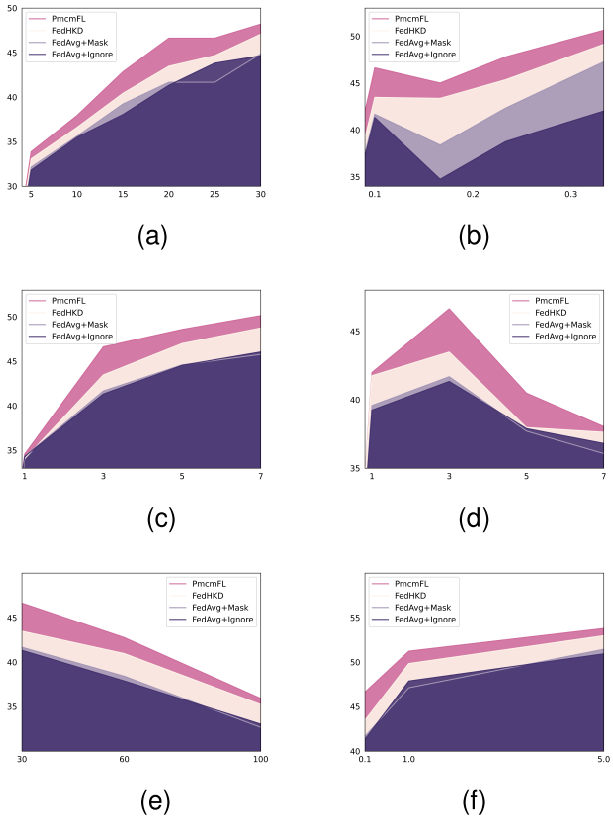


Fig. 7. Performance comparison of PmcmFL and baselines under different FL settings. Our PmcmFL consistently achieves the best performance. (a) Communication rounds. (b) Client selection rates. (c) Local training epochs. (d) Learning rates (10^{-5}). (e) Client counts. (f) Non-IID heterogeneity α.

In our PmcmFL framework, the parameters transmitted during each federated communication round include both the model weights and the prototype library. The number of parameters required for the prototype library increases linearly with the number of categories. Table VI illustrates different FL frameworks' communication overhead. When employing our tinyVQA, PmcmFL introduces merely a negligible increase in communication overhead (+ 0.32%). Even with the original VQAv2 dataset, PmcmFL demonstrates lower communication overhead compared to FedHKD (+ 3.21% versus + 5.43%). This underscores the superiority of our method in managing communication overhead. Furthermore, as shown in Fig. 6, we can also see that our PmcmFL tends to get better performance when passing the same amount of parameters (or rather, it requires less total communication overhead to achieve the same performance). This shows that our method has the highest communication efficiency.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

TABLE VII
PERFORMANCE OF THE GLOBAL MODEL UNDER DIFFERENT PROTOTYPE COMPUTATION METHODS. THE OPTIMIZED PROTOTYPE CONSTRUCTION METHOD ELIMINATES OUTLIERS

Definition	Accuracy (%) on 5K Testing Samples				
	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$
Original	53.810	50.606	44.916	41.014	38.264
Optimized	**54.215**	**50.793**	**45.448**	**41.240**	**38.541**

H. Robustness to Federated Settings

Robustness to different federation settings determines the scalability of an FL framework. Therefore, we compare our PmcmFL to representative baselines under various federation settings. The federation hyperparameters considered include communication rounds, client selection rates, local training epochs, learning rates, client counts, and the non-IID heterogeneity hyperparameter α. We adjust these federation hyperparameters to values different from the implementation details while keeping the other hyperparameters unchanged. Without loss of generality, we set the modality missingness rate to 0.5 for the training phase and use modality-complete test data for the testing phase. Fig. 7 illustrates the performance of various FL Frameworks. We compare four representative FL frameworks and corresponding missingness handling strategies: 1) FedAvg with Ignoring Missingness; 2) FedAvg with Noise Mask; 3) FedHKD with Noise Mask; 4) our PmcmFL. The results demonstrate that our PmcmFL consistently achieved optimal performance under different FL settings, showing robustness across various FL settings.

I. Optimizing Prototype Definition

We calculated the Z-score for each data point and treated the top 5% with the highest Z-scores as outliers to be removed. Based on this, we conducted experiments. The results are shown in Table VII. Under different modality-missing rates, the optimized prototype computation method leads to improved performance of the global model.

J. Discussion on the Security of Prototypes

Compared to traditional FL, PmcmFL introduces an additional prototype library, the security of which warrants consideration. On one hand, prototypes are computed by averaging all data points of the same class, which significantly smooths out the distinctive characteristics of individual data samples. As a result, the prototypes carry minimal risk of privacy leakage. Moreover, in our setting, the prototype library accounts for only 0.32% of the model parameters, making it highly compact. This compactness means it contains very few client-specific attributes, and under inference-based adversarial attacks, the prototype library reveals almost no client attribute information. On the other hand, the prototype library demonstrates strong resilience against model poisoning attacks. In our setup with 30 clients and a 0.3 modality-missing rate during training, we designated one multimodal client to carry out an attack by uploading random local prototypes. The final accuracy of the global model under this scenario is 51.327% (only 0.929% performance drop), indicating that our method is capable of defending against poisoning attacks.

It is worth noting that while PmcmFL can resist adversarial attacks, where an attacker cannot precisely reconstruct the client's original data from prototypes. However, the prototypes still provide a simple overview of each client's data distribution, which poses a potential privacy leakage risk.

VI. CONCLUSION

In this article, we investigate mFL with uncertain modality missingness in both training and testing. Our work introduces a prototype library to empower the FL framework with the capability to alleviate performance degradation resulting from modality missingness. Based on this prototype library, we construct a task-calibrated training loss, a model-agnostic unimodal inference strategy, and a proximal term. The effectiveness of our framework has been validated through comparisons with state-of-the-art methods under various modality missingness settings.

Currently, there are still certain limitations of the proposed PmcmFL framework in prototype computation; for example, we are unable to obtain consistent prototypes across datasets. Moreover, the effectiveness of the prototype mask during the inference phase is limited and requires further improvement. Therefore, our future work will primarily focus on enriching the semantics of prototypes, improving prototype matching accuracy, and exploring more representative and consistent prototypes, as well as leveraging prototypes to address missing or inaccurate labels. In addition, we have noted certain security concerns with the prototype library and plan to conduct research on protecting its privacy.

REFERENCES

[1] L. Zhen, P. Hu, X. Peng, R. S. M. Goh, and J. T. Zhou, "Deep multimodal transfer learning for cross-modal retrieval," in *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, pp. 798–810, 2020.
[2] W. Wang et al., "Image as a foreign language: BEIT pretraining for vision and vision-language tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19175–19186.
[3] J. Yu, Z. Wang, V. K. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "CoCa: Contrastive captioners are image-text foundation models," *Trans. Mach. Learn. Res.*, vol. 1, pp. 1–20, Aug. 2022.
[4] X. Chen et al., "PaLI: A jointly-scaled multilingual language-image model," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–33.
[5] D. Wang, T. Zhao, W. Yu, N. V. Chawla, and M. Jiang, "Deep multimodal complementarity learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10213–10224, Dec. 2023.
[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, vol. 54, A. Singh and J. Zhu, Eds., Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.
[7] L. Cao, H. Chen, X. Fan, J. Gama, Y.-S. Ong, and V. Kumar, "Bayesian federated learning: A survey," in *Proc. Int. Joint Conf. Artif. Intell.*, 2023, pp. 7233–7242.
[8] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3710–3722, Aug. 2021.
[9] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3400–3413, 2019.
[10] L. Cao, "Non-IID federated learning," *IEEE Intell. Syst.*, vol. 37, no. 2, pp. 14–15, Mar. 2022.
[11] Z. Ma et al., "FlGan: GAN-based unbiased federated learning under non-IID settings," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 4, pp. 1566–1581, Apr. 2024.
[12] W. Huang, M. Ye, and B. Du, "Learn from others and be yourself in heterogeneous federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10133–10143.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

[13] L. Qu et al., "Rethinking architecture design for tackling data heterogeneity in federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10051–10061.

[14] H. Jamali-Rad, M. Abdizadeh, and A. Singh, "Federated learning with taskonomy for non-IID data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8719–8730, Nov. 2023.

[15] Q. Yu, Y. Liu, Y. Wang, K. Xu, and J. Liu, "Multimodal federated learning via contrastive representation ensemble," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–15.

[16] Y. Zhao, P. Barnaghi, and H. Haddadi, "Multimodal federated learning on IoT data," in *Proc. IEEE/ACM 7th Int. Conf. Internet-of-Things Design Implement. (IoTDI)*, May 2022, pp. 43–54.

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

[18] T. Feng et al., "FedMultimodal: A benchmark for multimodal federated learning," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2023, pp. 4035–4045.

[19] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "BEiT v2: Masked image modeling with vector-quantized visual tokenizers," 2022, *arXiv:2208.06366*.

[20] H. Bao, D. Li, F. Wei, and W. Furu, "BEiT: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–18.

[21] L. Zhou, M. Ye, D. Zhang, C. Zhu, and L. Ji, "Prototype-based multisource domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5308–5320, Oct. 2022.

[22] X. Liao et al., "HyperFed: Hyperbolic prototypes exploration with consistent aggregation for non-IID data in federated learning," in *Proc. 32nd Int. Joint Conf. Artif. Intell.*, Aug. 2023, pp. 3957–3965.

[23] W. Huang, M. Ye, Z. Shi, H. Li, and B. Du, "Rethinking federated learning with domain shift: A prototype view," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16312–16322.

[24] Y. Dai, Z. Chen, J. Li, S. Heinecke, L. Sun, and R. Xu, "Tackling data heterogeneity in federated learning with class prototypes," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 7314–7322.

[25] J. Xu, X. Tong, and S. Huang, "Personalized federated learning with feature alignment and classifier collaboration," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–25.

[26] H. Chen, W. Johnny, and H. Vikalo, "The best of both worlds: Accurate global and personalized models through federated learning with data-free hyper-knowledge distillation," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–24.

[27] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[28] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5583–5594.

[29] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 34892–34916.

[30] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "SMIL: Multimodal learning with severely missing modality," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2302–2310.

[31] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2021, pp. 2608–2618.

[32] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal learning with missing modality via shared-specific feature modelling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15878–15887.

[33] H. Wang, S. Luo, G. Hu, and J. Zhang, "Gradient-guided modality decoupling for missing-modality robustness," 2024, *arXiv:2402.16318*.

[34] Z. Pan, S. Jiang, X. Yang, H. Yuan, and J. Wang, "Hierarchical cross-modal image generation for multimodal biometric recognition with missing modality," *IEEE Trans. Inf. Forensics Security*, vol. 20, pp. 4308–4321, 2025.

[35] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18156–18165.

[36] W. Yao, K. Yin, W. K. Cheung, J. Liu, and J. Qin, "DrFuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 16416–16424.

[37] J. Kim, H. Kang, S. Kim, K. Kim, and C. Park, "Disentangling and generating modalities for recommendation in missing modality scenarios," in *Proc. 48th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2025, pp. 1820–1829.

[38] R. Wu, H. Wang, F. Dayoub, and H. Chen, "Segment beyond view: Handling partially missing modality for audio-visual semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 6100–6108.

[39] Z. Cheng, J. Lang, T. Zhong, and F. Zhou, "Seeing the unseen in micro-video popularity prediction: Self-correlation retrieval for missing modality generation," in *Proc. 31st ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Jul. 2025, pp. 142–152.

[40] H. Chen, H. Liu, L. Cao, and T. Zhang, "Bayesian personalized federated learning with shared and personalized uncertainty representations," 2023, *arXiv:2309.15499*.

[41] J. Snell, K. Swersky, and R. S. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 30, 2017, pp. 4077–4087.

[42] S. Nadal, A. Abelló, O. Romero, S. Vansummeren, and P. Vassiliadis, "Graph-driven federated data management," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 509–520, Jan. 2023.

[43] J. Li, F. Li, L. Zhu, H. Cui, and J. Li, "Prototype-guided knowledge transfer for federated unsupervised cross-modal hashing," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 1013–1022.

[44] B. Li, W. Gao, J. Xie, M. Gong, L. Wang, and H. Li, "Prototype-based decentralized federated learning for the heterogeneous time-varying IoT systems," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 6916–6927, Feb. 2024.

[45] E. Enouen, S. Caldas, M. Goswami, and A. Dubrawski, "PICSR: Prototype-informed cross-silo router for federated learning (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 23482–23483.

[46] J. Zhang, Y. Liu, Y. Hua, and J. Cao, "FedTGP: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning," 2024, *arXiv:2401.03230*.

[47] G. Wan, W. Huang, and M. Ye, "Federated graph learning under domain shift with generalizable prototypes," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 15429–15437.

[48] Y. Tan et al., "FedProto: Federated prototype learning across heterogeneous clients," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 8432–8440.

[49] X. Yu, X. Wu, X. Wang, D. Li, M. Yang, and P. Cheng, "FedSaaS: Class-consistency federated semantic segmentation via global prototype supervision and local adversarial harmonization," in *Proc. 33rd Int. Joint Conf. Artif. Intell.*, Aug. 2024, pp. 6922–6930.

[50] L. Fu, S. Huang, Y. Lai, T. Liao, C. Zhang, and C. Chen, "Beyond federated prototype learning: Learnable semantic anchors with hyperspherical contrast for domain-skewed data," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, 2025, pp. 16648–16656.

[51] K. Jin, L. Xu, X. Wang, S.-Y. Hsieh, J. Wu, and L. Lin, "FedCPD: Personalized federated learning with prototype-enhanced representation and memory distillation," in *Proc. 33rd Int. Joint Conf. Artif. Intell.*, Aug. 2024, pp. 5498–5507.

[52] S. Guo, H. Wang, S. Lin, Z. Kou, and X. Geng, "Addressing skewed heterogeneity via federated prototype rectification with personalization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 5, pp. 8442–8454, May 2025.

[53] Y. Zhou et al., "FedSA: A unified representation learning via semantic anchors for prototype-based federated learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 21, pp. 23009–23017, Apr. 2025.

[54] W. Huang, D. Wang, X. Ouyang, J. Wan, J. Liu, and T. Li, "Multimodal federated learning: Concept, methods, applications and future directions," *Inf. Fusion*, vol. 112, Dec. 2024, Art. no. 102576.

[55] M. Tan et al., "FedSea: Federated learning via selective feature alignment for non-IID multimodal data," *IEEE Trans. Multimedia*, vol. 26, pp. 5807–5822, 2024.

[56] Z. Wu, Q. Zhang, D. Miao, K. Yi, W. Fan, and L. Hu, "HyDiscGAN: A hybrid distributed cGAN for audio-visual privacy preservation in multimodal sentiment analysis," in *Proc. Int. Joint Conf. Artif. Intell.*, 2024, pp. 6550–6558.

[57] Q. Dai, D. Wei, H. Liu, J. Sun, L. Wang, and Y. Zheng, "Federated modality-specific encoders and multimodal anchors for personalized brain tumor segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 1445–1453.

[58] Z. Wu et al., "Seeking personality yet commonality via adversarial learning to enhance heterogeneous multimodal collaboration," *Inf. Fusion*, vol. 126, Feb. 2026, Art. no. 103578.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.

[59] J. Chen and A. Zhang, "Fedmbridge: Bridgeable multimodal federated learning," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 7667–7686.
[60] M. Gao, H. Zheng, X. Feng, and R. Tao, "Multimodal fusion using multi-view domains for data heterogeneity in federated learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, 2025, pp. 16736–16744.
[61] P. Saha, D. Mishra, F. Wagner, K. Kamnitsas, and J. A. Noble, "Incongruent multimodal federated learning for medical vision and language-based multi-label disease detection," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 27, pp. 28331–28339, Apr. 2025.
[62] H. Song, J. Wang, J. Zhou, and L. Wang, "Tackling modality-heterogeneous client drift holistically for heterogeneous multi-modal federated learning," *IEEE Trans. Med. Imag.*, vol. 44, no. 4, pp. 1931–1941, Apr. 2025.
[63] Z. Zhang, F. Qi, and C. Xu, "Enhancing storage and computational efficiency in federated multimodal learning for large-scale models," in *Proc. Int. Conf. Mach. Learn.*, 2024, pp. 59685–59699.
[64] L. Tran, W. Sun, S. Patterson, and M. Milanova, "Privacy-preserving personalized federated prompt learning for multimodal large language models," in *Proc. Int. Conf. Learn. Represent.*, 2025, pp. 1–19.
[65] C. Feng, D. Feng, G. Huang, Z. Liu, Z. Wang, and X.-G. Xia, "Robust privacy-preserving recommendation systems driven by multimodal federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 5, pp. 8896–8910, May 2025.
[66] R. Rahman, F. Patadia, J. Wang, and D. C. Nguyen, "Multimodal federated learning for air quality estimation with model personalization over aerial–ground networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 22, pp. 1–5, 2025.
[67] H. Q. Le et al., "Cross-modal prototype based multimodal federated learning under severely missing modality," *Inf. Fusion*, vol. 122, Oct. 2025, Art. no. 103219.
[68] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–21.
[69] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MI, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.
[70] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.
[71] H. Zhang, M. Cissé, Y. Dauphin, and D. López-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.
[72] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6325–6334.
[73] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[74] T.-M. Harry Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," 2019, *arXiv:1909.06335*.
[75] L. Tian, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, 2018, pp. 429–450.
[76] L. v. d. Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

Guangyin Bao received the B.E. degree from Tongji University, Shanghai, China, in 2022, where he is currently pursuing the M.S. degree with the College of Electronic and Information Engineering.

His current research interests focus on multimodal learning and federated learning.

Qi Zhang received the dual Ph.D. degree from Beijing Institute of Technology, Beijing, China, in 2020, and the University of Technology Sydney, Sydney, NSW, Australia, in 2022.

He is currently an Associate Professor at Tongji University, Shanghai, China. He has published high-quality papers in premier conferences and journals, including Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML), International Conference on Learning Representations (ICLR), AAAI Conference on Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI), ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), The Web Conference (TheWebConf), IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS. His primary research interests focus on collaborative filtering, sequential recommendation, learning to hash, and MTS analysis.

Duoqian Miao is currently a Professor and a Ph.D. Tutor with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. He has published more than 200 papers in IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON FUZZY SYSTEMS, *Pattern Recognition*, *Information Sciences*, and *Knowledge-Based Systems*. His research interests include machine learning, data mining, big data analysis, granular computing, artificial intelligence, and text image processing.

Dr. Miao's representative awards include the Second Prize of Wuwenjun AI Science and Technology in 2018, the First Prize of Natural Science of Chongqing in 2010, the First Prize of Technical Invention of Shanghai in 2009, and the First Prize of Ministry of Education Science and Technology Progress Award in 2007. He serves as the President for the International Rough Set Society (IRSS), the Chair for the CAAI Granular Computing Knowledge Discovery Technical Committee, the Vice Director for the MOE Key Laboratory of Embedded System and Service Computing, and the Vice President for Shanghai Computer Federation and Shanghai Association for Artificial Intelligence. He serves as an Associate Editor for *International Journal of Approximate Reasoning* and *Information Sciences* and an Editor of *Journal of Computer Research and Development* (in Chinese).

Zixuan Gong received the B.E. and M.S. degrees from Tongji University, Shanghai, China, in 2021 and 2025, respectively.

He has published high-quality papers in premier conferences and journals, including Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML), International Conference on Learning Representations (ICLR), AAAI Conference on Artificial Intelligence (AAAI), ACM Microwave Magazine, and *Transactions on Machine Learning Research* (TMLR). His current research interests focus on multimodal learning and neuroscience.

Chaochao Chen (Senior Member IEEE) received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2016.

He was a Visiting Scholar with the University of Illinois Urbana-Champaign, Illinois, IL, USA, from 2014 to 2015. He is currently a Distinguished Research Fellow at Zhejiang University. He has published more than 100 papers in peer-reviewed journals and conferences. His research mainly focuses on recommender systems, privacy-preserving machine learning, and graph machine learning.

Liang Hu received the dual Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2015, and the University of Technology Sydney, Sydney, NSW, Australia, in 2018.

He is currently a Distinguished Research Fellow at Tongji University, Shanghai, and a Chief AI Scientist at the DeepBlue Academy of Sciences, Shanghai. His research interests include recommender systems, machine learning, data science, and general intelligence.

Longbing Cao received the Ph.D. degree in pattern recognition and intelligent systems from Chinese Academy of Sciences, Beijing, China, and the Ph.D. degree in computing sciences from the University of Technology Sydney, Sydney, NSW, Australia.

He is a Distinguished Chair Professor at Macquarie University, Sydney. His research interests include artificial intelligence, data science, machine learning, and behavior informatics.

Dr. Cao is an ARC Future Fellow (professorial level) and the Editor-in-Chief of IEEE INTELLIGENT SYSTEMS and *International Journal of Data Science and Analytics*.

Authorized licensed use limited to: TONGJI UNIVERSITY. Downloaded on November 11,2026 at 14:03:06 UTC from IEEE Xplore. Restrictions apply.