A Contrastive Learning and Graph-based Approach for Missing Modalities in Multimodal Federated Learning

Thu Hang Phung*, Binh P. Nguyen‡, Thanh Hung Nguyen*§,
Quoc Viet Hung Nguyen†, Phi Le Nguyen* and Thanh Trung Huynh¶

*School of Information and Communication Technology, Hanoi University of Science and Technology, Vietnam
Email: Hang.PT232173M@sis.hust.edu.vn, hungnt,lenp@soict.hust.edu.vn
†School of Information and Communication Technology, Griffith University, Australia
Email: quocviethung.nguyen@griffith.edu.au
‡School of Mathematics and Statistics, Victoria University of Wellington, New Zealand
Email: binh.p.nguyen@vuw.ac.nz
¶Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland
Email: thanh.huynh@epfl.ch

***Abstract*—Federated Learning has emerged as a decentralized method for training machine learning models using distributed data sources. It ensures privacy by allowing clients to collaboratively learn a shared global model while keeping their data stored locally. However, a significant challenge arises when dealing with missing modalities in clients' datasets, where certain features or modalities are unavailable or incomplete, leading to heterogeneous data distribution. Previous studies have addressed this issue, but they fall short in addressing generalizability across diverse, unobserved individuals. This study introduces MIFL, a novel framework for handling modality-missing clients in Multimodal Federated Learning. MIFL utilizes a unique approach, optimizing a client's local model with existing modalities while incorporating absent modalities from clients. Aggregation of these models is performed through a graph-based attentive aggregation method, maintaining generalized characteristics by updating a global model averaged across clients. Our experimental results demonstrate the effectiveness of MIFL across various client configurations with statistical heterogeneity, showcasing its potential for addressing the challenge of missing modalities in Federated Learning.**

***Index Terms*—Federated learning, multimodal learning, modality missing, graph-based attentive aggregation**

I. Introduction

Multimodal learning, a novel machine learning paradigm, has surfaced as a promising approach for decision-making using diverse input modalities. Unlike traditional methods that rely on a specific modality (such as image, sound, or text), multimodal learning integrates various modalities, offering distinct perspectives on the same event. This integration results in an enriched view, contributing to increased accuracy [1]. Multimodal learning has garnered considerable attention due to its potential, leading to numerous proposed studies. Initially, research in this field concentrated on developing model architectures to address modality fusion and translation [1]–[4]. However, these studies commonly assumed that both training and testing data possessed complete modalities, which is impractical. Indeed, empirical evidence has shown that, in practice, certain data pieces often lack specific modalities [1], [5]. In response, substantial recent efforts have been directed toward addressing challenges associated with missing modalities, which can be categorized into two groups: methods that incorporate data imputation [6], [7] and non-imputation mehods [8]–[10]. Concerning the first group, Zhang et al. [6] reconstruct missing data in a modality by calculating the average of values from similar modality samples, which does not take into account the correlation between different modalities. Zhou et al. [7] introduce a conditional generator for brain tumor segmentation that utilizes available modalities to generate the information for the absent ones. Regarding the second group, solutions have been proposed so that models can benefit from either complete samples through contrastive-aid learning [10] or only the available missing modalities by optimal sparse linear prediction [8] or by hypergraph-based fusion [9]. However, all the aforementioned studies focus on centralized learning, which entails gathering and training all data in a single location. This centralized approach has demonstrated several drawbacks, with the most severe being privacy compromises [11].

To address privacy concerns, Google introduced Federated Learning (FL) in 2017 [12], offering a novel distributed training mechanism. This approach enables clients (data owners) to collaboratively train a model under the supervision of a central server while keeping their data locally, effectively mitigating privacy leakage. Federated Learning involves two primary components: a local training process on the client side and aggregation on the server side. In this process, clients utilize their data to locally train the model and then transmit

§Corresponding author

its weights to the server for aggregation, creating a global model. Early versions of Federated Learning [13]–[20] were designed exclusively for unimodal learning, where all clients' data shared the same modality. Recognizing the significance of multimodal learning, subsequent studies have proposed Federated Learning mechanisms for multimodal scenarios [21]–[23]. However, most of these solutions only address the issue when all clients possess data with complete modalities. The necessity for complete modalities, once deemed impractical in centralized learning, becomes even more challenging in the context of Federated Learning due to the heterogeneity of clients' data. Consequently, addressing missing modalities in multimodal FL is a crucial challenge. Unfortunately, this problem has not received extensive attention in the literature, with only a few existing works tackling it. Chen et al. [24] propose a framework for personalized FL, thereby not capable of producing a global model generalizing over the private testing dataset. The approach developed by Yu et al. [11] typically demands the server possess a dataset, which is often impractical in most FL scenarios.

This study specifically addresses the challenge of solving the missing modality problem in multimodal FL. We focus on scenarios where each client's data may lack certain modalities, and the sets of modalities held by clients are not uniform. In Fig. 1, we illustrate an example depicting the degradation of a FedAvg in the context of missing modalities. Specifically, we conduct the experiment on a 12-modality dataset [25] with the FL setting of clients possessing varying numbers of modalities: 2 to 6 modalities for the blue lines, while the red line represents clients with all 12 modalities. The significant disparities between testing including all modalities and testing involving only a few modalities serve as empirical data supporting the necessity of enhancing modality-missing FL performance. The challenge of the modality missing problem in multimodal FL arises from two perspectives. Firstly, in the traditional FL model, clients' data shares the same modality, allowing the use of a uniform model structure for all clients. However, in this new context, each client possesses a distinct set of modalities, or, in other words, different data types. This raises a research question: *How should we design the local model to handle data with this varying set of modalities?* Secondly, considering that each local model of a client learns distinct knowledge from a unique set of modalities, a crucial question arises: *How can we design an aggregation mechanism on the server side that effectively combines models from all clients?*

This study introduces a novel multimodal FL mechanism named MIFL (stands for **M**odality-m**I**ssing **F**ederated **L**earning), capable of simultaneously addressing the challenges outlined above. Our approach effectively addresses missing modalities in both training and test data. Importantly, it does not require the server to possess any data, providing a robust and privacy-conscious solution. To tackle the heterogeneous modality problem, we propose an extra-modality generator, generating an additional feature for each modality. These extra features are then employed to fill in the corresponding missing modality, allowing synchronization of each client's modality set and facilitating the design of local models for clients. Furthermore, we incorporate a contrastive learning mechanism that aligns features generated by partial-modality data with those from full-modality data. This approach allows our model to enrich the information derived from modality-missing data, thereby enhancing the effectiveness of the multimodal FL framework.

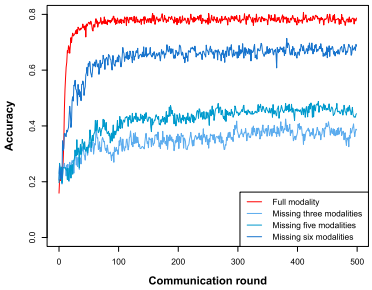

Fig. 1: Performance discrepancy of FedAvg with missing modality clients.

The main contributions of the research are as follows.

- We present a novel Federated Learning mechanism designed to tackle the multimodality learning problem, adept at addressing the challenge of missing modalities during both training and inference stages.
- To overcome the issue of modality heterogeneity on the client side, we introduce an innovative architecture for the local model, featuring the introduction of an extra-modality generator. This generator plays a crucial role in producing additional representation vectors, each corresponding to a specific modality. These modality representation vectors are employed to substitute for missing modalities.
- Leveraging the contrastive learning paradigm, we design an auxiliary loss that aligns the representation of data lacking certain modalities with those possessing complete modalities. This loss function enables the enrichment of information extracted from data lacking modalities, thereby enhancing the model's accuracy.
- We propose a server-side aggregation mechanism designed to effectively aggregate models from clients with diverse methodologies.

The remainder of this paper is organized as follows. Section II presents the specifics of our proposed algorithm for multimodality-missing Federated Learning. The experimental results are explained in Section III. Finally, the conclusion and potential future directions are discussed in Section IV.

II. Proposed Method

In the following, we delve into the details of our proposed method. Initially, we introduce the problem formulation and the motivation behind our solution in Section II-A. Section II-B offers an overview of our proposal, while Sections II-C and II-D provide detailed information.

A. Problem Formulation and Motivation

Problem Formulation. Consider a FL system that consists of K clients ($K \geq 2$), represented by $C_1, ..., C_K$. We denote by D_i the training dataset of client D_i, whose cardinality is depicted by n_i ($i = 1, ..., K$). Let us denote by $\mathcal{M}$ the set of all modalities and M_i the set of modalities possessed by client C_i ($M_i \subseteq \mathcal{M}$). The goal of our FL model is to determine a global model ω satisfying the following objective function:

$$\omega = \min_{\omega} \frac{1}{\sum_i n_i} \sum_{i=1}^{K} \sum_{j=1}^{n_i} f_i \left(\omega, X_i^j, y_i^j\right), \tag{1}$$

where (X_i^j, y_i^j) is the j-th sample of D_i, and $f_i \left(\omega, X_i^j, y_i^j\right)$ is the loss function of D_i's local model. In this work, we focus on the classification task.

B. Overview of the Framework

We propose an innovative framework for Multimodal Federated Learning, featuring a modality-aware and generalized client's model combined through a graph-based aggregation method and updated with contrastive-aid training (Figure 2). To overcome heterogeneous clients, we propose a modality-specific and generalized local model for each client. Its training process employs contrastive learning to enhance feature extractor-generated representations. Additionally, an attentive graph-based aggregation method is implemented to leverage the similarities among clients' datasets. Our study pioneers efforts to enhance the generalization capabilities of Multimodal Federated Learning while addressing challenges in non-IID data settings among clients.

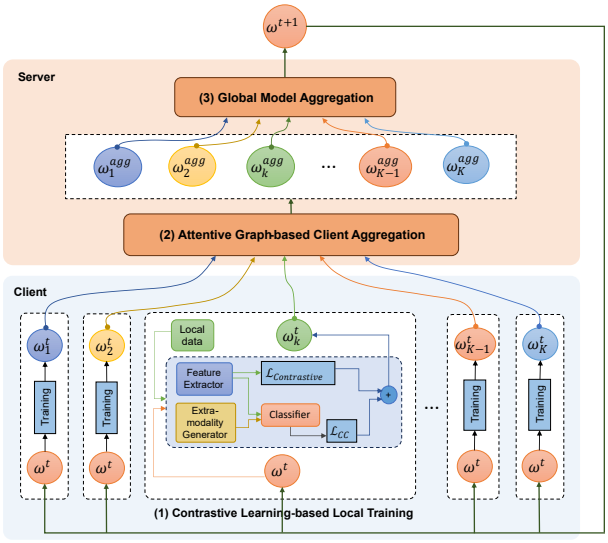

Fig. 2: Overview of the proposed framework.

C. Modality-aware Local Model for Clients

Figure 3 illustrates the architecture of our proposed local model for clients. The details of the first two elements are described in subsections II-C1 and II-C2, while the classifier is just a simple Fully Connected Layer. The additional information and modality embeddings, which are outputs from the feature extractor, are combined using the element-wise sum operator after being generated. To create the optimal hidden vector before inputting it into the classifier, the embeddings from the existing modalities are joined with the embeddings generated for the missing modalities using concatenation.

1) Feature Extractor: Let d_H be the embedding dimension outputted from each feature extractor. For a client k, given its input dataset $\mathcal{D}_k$, the output embedding h_{kj}^i of sample i belonging to this client's feature extractor of the existing modality indexed by j, X_{kj}^i, is obtained as follows:

$$h_{kj}^i = \underset{w_{kj}}{\text{FeatureExtractor}}(X_{kj}^i) \in \mathbb{R}^{d_H}, \tag{2}$$

where w_{kj} is the weight of the client k's feature extractor in modality j.

2) Extra Modality Information Generator: To address the issue of missing modalities in FL systems, most existing methods include aggregating similar modalities from all clients. However, this technique restricts the local model's capabilities, limiting it to only the modalities it currently possesses. This arrangement thereby imposes a limitation on accuracy, as the highest degree of performance that each individual client can achieve depends on the specific type of data it has. Furthermore, it does not make use of the potential advantages that could be obtained by integrating information from diverse modalities across multiple clients.

This phenomenon has motivated our proposal to integrate an additional information generator corresponding to each modality. The additional information will be converted into weight vectors and is expected to demonstrate modality generalizability across different clients. In particular, for each modality, there are two possibilities: it can either exist or be absent. Hence, we propose implementing a weight matrix that can be updated and consists of two rows, where each row represents one of the two scenarios of modality presence. The selection of a suitable vector from two potential rows for the additional modality data relies on the presence of each modality in each client. The selection is specified using a binary vector consisting of two items. If modality j exists in the training dataset of a client, its one-hot vector will be $[1, 0]$; otherwise, it will be $[0, 1]$. The first vector represents the additional modality information in cases where the client already has that modality, whereas the other vector can be seen as the generative output of the missing modality.

Figure 4 demonstrates that the presence of each modality is considered when determining which row vector in the weight matrix will be generated as output. This is a learnable weight matrix that will be updated during the training process of the federated system. Each client will have M weight matrices corresponding to M modalities. Let $\widetilde{\mathcal{W}}_j$ be the weight matrix of modality j. Specifically, if client k possesses modality j, its additional data for modality j is represented as $\widetilde{h_{kj}} = [1, 0] \times \widetilde{\mathcal{W}}_j \in \mathbb{R}^{d_H}$; otherwise, $\widetilde{h_{kj}} = [0, 1] \times \widetilde{\mathcal{W}}_j \in \mathbb{R}^{d_H}$.

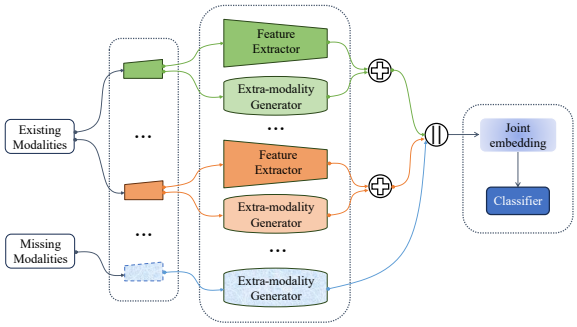

Fig. 3: Detailed architecture of the local model.

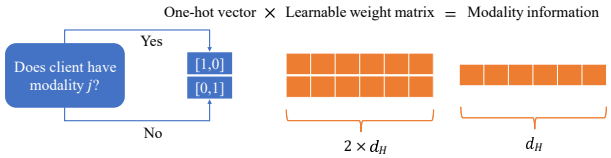

Fig. 4: Extra modality information generation process.

Let $h_{kj} \in \mathbb{R}^{d_H}$ be the output of modality j's feature extractor for client k, the final joint embedding $h_k \in \mathbb{R}^{d_H}$ is calculated as follows:

$$h_k = \left(\Big\| _{i \in \mathcal{M}_k} (h_{ki} \oplus \widetilde{h_{kj}}) \right) \left(\Big\| _{i \notin \mathcal{M}_k} \widetilde{h_{kj}} \right), \tag{3}$$

where $\|$ and $\oplus$ are the concatenation and element-wise sum operators, respectively. From this equation, we can derive that $h_k \in \mathbb{R}^{d_H \times M}$.

3) Contrastive Learning-based Local Training: This study employs contrastive learning to enhance the alignment between uni-modal embeddings, which are representations acquired by individual feature extractors, and the combined representation derived by element-wise summation of these uni-modal embeddings.

Denote the proposed multimodal contrastive loss by MultimodalCL. Consider the outputs of client k's feature extractor (uni-modal representations) $h_{kj}, j \in \mathcal{M}_k$ and a mini-batch $\mathcal{B}$ of size B. Let $\widehat{h_k} = \{\oplus h_{kj}, j \in \mathcal{M}_j\}$ be the multimodal representation obtained by applying element-wise sum on uni-modal embeddings.

Let $sim(u, v)$ denote the cosine similarity among vectors u and v. Denote by

$$s_{kj}(i,l) = \exp\left(sim\left(h_{kj}^i, \widehat{h_k^l}\right)\right) + \exp\left(sim\left(h_{kj}^l, \widehat{h_k^i}\right)\right) \tag{4}$$

the similarity between representations h_{kj} and $\widehat{h_k}$ of 2 samples i and l.

For a given modality j, we define positive pairs as $\left(h_{kj}^i, \widehat{h_k^l}\right)$ for $y^i = y^l$ and $i = 1, ..., B$. The remaining pairs are considered negative pairs. Then the described loss function with respect to mini-batch $\mathcal{B}$ can be expressed as:

$$\text{MultimodalCL}(\mathcal{B}) = \frac{1}{|\mathcal{M}_k| \times |\mathcal{B}|} \sum_{j=1}^{\mathcal{M}_k} \sum_{i=1}^{B} -\log \frac{s_{kj}(i,l)}{\sum_{y^i \neq y^l} s_{kj}(i,l)}. \tag{5}$$

This multimodal contrastive loss is added when training the client's local model. In particular, each client k independently performs local Mini-batch Gradient Descent for batch $\mathcal{B}$:

$$\ell(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} f_k\left(w, X^i, y^i\right). \tag{6}$$

Then, the final loss function for mini-batch $\mathcal{B}$ is calculated as follows:

$$\mathcal{L}(\mathcal{B}) = \ell(\mathcal{B}) + \alpha \text{MultimodalCL}(\mathcal{B}), \tag{7}$$

where α is a hyperparameter that determines the balance between the two losses.

This client updates its model's parameters for the mini-batch $\mathcal{B}$ following the below equation, meanwhile, it considers extra regularization term:

$$w_k \leftarrow w_k - \eta \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left(\nabla_{w_k} \mathcal{L}(\mathcal{B}) + \nabla_{w_k} \lambda R_k\left(\widehat{w_k}\right)\right), \tag{8}$$

where the multimodal relational information is achieved by minimizing the Mean Squared Error loss between the model and the approximation of neighborhood information $R_k(\widehat{w_k})$, which will be discussed in section II-D. η is the learning rate, and λ is a hyperparameter to balance global correlation and local personalization.

D. Attentive Graph-based Client Aggregation

1) Feature Extractor and Classifier Aggregation: We propose to implement a neighborhood-attentive graph aggregation method on the server to process the local model. This methodology draws inspiration from the FedMSplit architecture [24]. This aggregation strategy is specifically implemented in classifiers and feature extractors. The parameter vector at client k is denoted by w_k. The vector can be divided into the following components:

$$w_k = \{w_{ki} \mid \forall i \in \mathcal{M}_k\} \cup \{\widetilde{w}_{kj} \mid \forall j \in \mathcal{M}\} \cup \{\overline{w_k}\}, \tag{9}$$

where w_{ki}, $\widetilde{w}_{kj}$, and $\overline{w_k}$ are the weights of the feature extractor, additional information generator, and classifier, which correspond to the client, respectively.

The feature extractor and classifier weights $\widehat{w_k}$, shown in Figure 2 are comprised of

$$\widehat{w_k} = \{w_{ki} \mid \forall i \in \mathcal{M}_k\} \cup \{\overline{w_k}\}. \tag{10}$$

Following the traditional FL framework, a set of clients $\mathcal{S}_t$ take part in communication round t. To support heterogeneous and complex multi-space interactions, the FedMSplit method uses a two-step technique. Initially, it uses a multi-view, attentive, and graph-based message forwarding technique to estimate each client's complex neighborhood information by combining models from other clients. Following that, each client trains their model independently using a local Stochastic Gradient Descent on the dataset D_k.

The dynamic multi-view graph is denoted by $\mathcal{G}_t = (\mathcal{V}, \Phi_t, \mathcal{E}, \mathcal{A}_t)$. Details of the components are as follows:

- $\mathcal{V} = \{v_k\}_{k \in S_t}$ is the vertex set, each vertex v_k corresponds to a client k in S_t that contains a local multimodal dataset $D_k = \{(x^i, y^i)\}_{i=1}^{n_k}$.
- $\Phi_t = \{w_k^t\}_{k \in S_t}$ is the node's features, representing the model parameters of each client at round t. Those parameters are updated and change over time, so this graph is a dynamic graph.
- $\mathcal{E}$ is the edge set. This graph is configured to be a fully connected graph. However, a node's information cannot be transmitted completely and directedly due to the fact that each node's features are associated with distinct combinations of modalities.
- $\mathcal{A}_t \in \mathbb{R}^{|S_t| \times |S_t| \times (M+1)}$ represent the edge features. An edge feature $\Lambda_{k,l}(t)$ indicates the model weights similarities between two clients and consists of *multiple dimensions* (multi-view) of Euclidean distances; each dimension corresponds to a type of block in client models. If client *k* and client *l* do not have the common block *j*, $\Lambda_{k,l,j}(t) = 0$.

Among these components, $\mathcal{A}_t$ and Φ_t change in each communication round. Updating these terms can be performed in sequential order. In each communication round t, updating correlation $\Lambda(t)$ fixing Φ_t is treated as updating edge features based on the current node's features (client embeddings). Formally, for each pair of clients (*k*, *l*) and each model block *j*, their relationship can be measured as $\Omega_{j,kl} = Att(w_{kj}, w_{lj})$ using any metric or attention function $Att(\cdot, \cdot)$, such as additive attention, dot product, and multiplicative attention. In the experiments, dot product attention is utilized for all simulations. This aggregation can be described as follows:

$$\begin{cases} \overline{w}_k^{agg} \leftarrow \sum_{l=1}^{\mathcal{N}} \frac{q(\Lambda_{kl})\bar{\Omega}_{kl}}{\sum_{p=1}^{\mathcal{N}} q(\Lambda_{kp\cdot})\bar{\Omega}_{kp}} \overline{w}_l, \\ w_{kj}^{agg} \leftarrow \sum_{l=1}^{\mathcal{N}} \frac{q(\Lambda_{kl\cdot})\Omega_{j,kl}}{\sum_{p=1}^{\mathcal{N}} q(\Lambda_{kp\cdot})\Omega_{j,kp}} w_{lj} \text{ for } \forall j \in \mathcal{M}_k \cap \mathcal{M}_l, \end{cases} \tag{11}$$

where $q(\Lambda_{kl}) = ||\Lambda_{k,l,\cdot}||_1/(1 + |\mathcal{M}_k \cap \mathcal{M}_l|)$, and $\mathcal{N}$ is the number of neighbors of client k. Since all selected clients are considered neighbors, $\mathcal{N} = S_t$.

2) Additional Modality Information Aggregation: Applying the Federated Averaging (FedAvg) algorithm on the set of M additional modality generators possessed by each client is straightforward. However, customizing this component to optimize individual client performance may be inconsistent with its intended purpose, as it is intended to be generalized across clients. Each client's weight matrix for a specific modality is updated by modifying only one row. This occurs because all samples belonging to the same client have the same combination of modalities, leading to the generation of a single one-hot vector for each instance. In order to minimize any possible adverse effects on the global model, the remaining row in the weight matrix is set to zeros.

Particularly, the global model's as well as the initial client model's extra information generator can be aggregated as follows:

$$\widetilde{w_i}^{agg} \leftarrow \sum_{k \in S_t} \frac{n_k}{n} \widetilde{w_{ki}}^{agg}, \forall i \in \mathcal{M}, \tag{12}$$

where $\widetilde{w}_{kir} = 0$ if the r-th row in the weight matrix of client k corresponds to its missing modality i.

At the end of the graph-based aggregation between clients, client k has the updated model denoted by:

$$w_k^{agg} = \left\{ \left\{ w_{kj}^{agg} \mid \forall j \in \mathcal{M}_k \right\} \cup \{\widetilde{w_{ki}}^{agg} \mid \forall i \in \mathcal{M}\} \cup \{\overline{w_k}^{agg}\} \right\}. \tag{13}$$

Then, updating model weight w_k^t can be viewed as updating node features (client models) based on local datasets as well as current edge features (client-client relationships). As mentioned before, each client k performs local Mini-batch Gradient Descent on batch $\mathcal{B}$, while applying an extra term:

$$w_k \leftarrow w_k - \eta \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left(\nabla_{w_k} \mathcal{L}(\mathcal{B}) + \nabla_{w_k} \lambda R_k \left(\widehat{w_k} \right) \right), \tag{14}$$

where $R_k(\widehat{w_k}) = \|\overline{w}_k - \overline{w}_k^{agg}\|_2^2 + \sum_{j \in \mathcal{M}_k} \|w_{kj} - w_{kj}^{agg}\|_2^2$.

This term has the effect of regularizing the local model so that it will not suffer from overfitting. Moreover, this term gives balance to the local model when regulating the aggregated one, which contains its attentive neighborhood information.

3) Global Model Aggregation: In our proposed framework, after aggregating all clients' models together, a global model denoted by w^t is obtained as follows:

$$w^t = \frac{1}{|S_t|} \sum_{k \in S_t} \{w_k^{agg}\}. \tag{15}$$

Averaging the global model allows for the evaluation of several modalities, ensuring robustness and adaptation in real-world settings involving different data sources. The global model's collaborative learning approach encourages an effective and inclusive learning process, lowering the impact of modality disparities throughout the network.

III. Experiments

A. Dataset

With annotations conforming to the SCP-ECG standard, the PTB-XL dataset [25], which consists of 21,837 clinical 12-lead electrocardiogram (ECG) recordings from 18,885 individuals, is a substantial resource, and it is used in our experiments. There are three distinct categories of annotations: form (significant changes in specific ECG segments), rhythm (specific changes in rhythm), and diag (diagnostic statements). The dataset has 71 distinct statements, including 44 diagnostic, 12 rhythm, and 19 form statements. Four form statements, in particular, function as diagnostic ECG statements. In addition, the diagnostic statements comprise 24 subclasses and five coarse superclasses, organized hierarchically. The largest publicly accessible clinical ECG dataset, this particular set is notable for its extensive annotations and metadata, which render it a valuable resource for training machine learning algorithms.

In this study, the PTB-XL dataset was modified to facilitate the process of data splitting and model evaluation. By ensuring that each sample is assigned to an exact single label, the subset rectifies the observation that diagnostic samples frequently

TABLE I: Data configurations

Method	Full-modal FedAvg	Missing-modal FedAvg	FedAvg	FedMSplit	MIFL
Training data	Full 12 modalities	Full 12 modalities	Experiment 1: 2-6 modalities		
			Experiment 2: 6-10 modalities		
Testing data	Full 12 modalities	Combination #1: 3 modalities [3, 7, 11]			
		Combination #2: 5 modalities [2, 3, 7, 11, 12]			
		Combination #3: 5 modalities [2, 3, 7, 10, 11]			
		Combination #4: 6 modalities [3, 5, 6, 10 ,11, 12]			
		Combination #5: 9 modalities [3, 4, 5, 6, 7, 8, 10, 11, 12]			
		Combination #6: 8 modalities [3, 5, 6, 7, 8, 9, 10, 12]			
		Combination #7: 10 modalities [1, 2, 3, 5, 6, 7, 8, 9, 10, 12]			

pertain to only one class. This reduced dataset consists of 3963 samples, which belong to five classes. This balanced dataset is subsequently divided into a training set and a test set in a 4:1 ratio.

B. Data Preparation

To assess the effectiveness of the proposed framework, we partitioned the refined dataset (with 12 modalities being 12 leads) into 20 distinct clients ($K = 20$). Introducing a more realistic multimodal missing setting compared to previous studies, these clients were assigned two distinct combinations of modality sets. We performed two experiments in which the number of modalities for each client was (1) a random count from 2 to 6 (*Experiment with small numbers of modalities*), and (2) a random count from 6 to 10 (*Experiment with large numbers of modalities*). It should be noted that each sample is exclusive to one client, and the original dataset comprises complete-modality samples. To establish the desired conditions, some modalities are manually excluded. The proposed framework is evaluated with seven testing scenarios outlined in Table I. Each combination varies in both the number and type of modalities. Each testing combination is specifically chosen to be less likely to appear in any of the three client settings. Determining the maximum number of existing times involves selecting the maximum occurrences of these testing combinations among the least-appearing modality combinations within a client.

C. Baselines

We evaluate the effectiveness of our proposed framework, MIFL, using FedAvg [26] and FedMSplit [24]. We also employed FedAvg to establish an upperbound for subsequent comparisons. Two baseline configurations utilizing FedAvg aggregations were incorporated into the experiment process:

- **Full-modal FedAvg**: All clients have complete observations with full modalities. A new testing set having the full-modal characteristics of the training set was generated for evaluation, eliminating the need for manual modality exclusion for each client.
- **Missing-modal FedAvg**: Similar to Full-modal FedAvg in the training configuration; however, this model was tested using the seven specified testing combinations instead of a full-modal training set.

All the aforementioned configurations are summarized in Table I.

D. Results

We set the number of client local training epochs $E = 3$ and used the following values for the hyperparameters of the models: number of communication rounds $T = 500$, percentage of selected clients at each round $C = 1.0$, batch size $B = 64$, learning rate $\eta = 0.5$, embedding dimension $d_H = 128$, 50 early stopping rounds, and FedMSplit's weight regularization $\lambda = 0.01$.

1) Accuracy Evaluation: Figure 5 shows the testing accuracy results for seven combinations in two client settings. It should be noted that the accuracy of Full-modal FedAvg is 0.807, which was reached only with the full-modality combination.

In both experiments, a recognizable trend develops, with the gap between MIFL and FedMSplit being most noticeable in combination #1 testing and gradually decreasing as the number of modalities in the testing set increases. FedAvg continually produces less accurate results due to its intrinsic simplicity, failing to appropriately handle the diverse needs of its clients. As the number of testing modalities increases, all three methods approach the accuracy values of Missing-modal FedAvg and Full-modal FedAvg, with MIFL being the most comparable, demonstrating its superiority.

In Experiment 2, the differences reduce and become closer to the two upper bounds, with significant variations seen in testing combinations #1 and #2. FedAvg achieves a high level of accuracy in most circumstances, coming quite close to the upper bounds. The progress of MIFL, especially considering its limited and varied number of modalities, is impressive.

MIFL regularly outperforms all other testing methods in terms of accuracy, demonstrating greater generality in a wide range of contexts. The proposed aggregated global model consistently achieves the highest precision values.

As can be seen in Tables II and III, MIFL improves from 1.64% to 48.41% when compared to the FedMSplit baseline, which can be attributed to different training scenarios and testing combinations. Improvements range from 6.4% to 48.41% for six or fewer training modalities and 1.64% to 9.81% in other training cases, demonstrating considerable progress. This figure is more significant when comparing MIFL and FedAvg, with an enhancement of 30.75% averaging over all testing combinations and experiments. The biggest improvement is 106.4%, which belongs to testing combination #1 of Experiment 2.

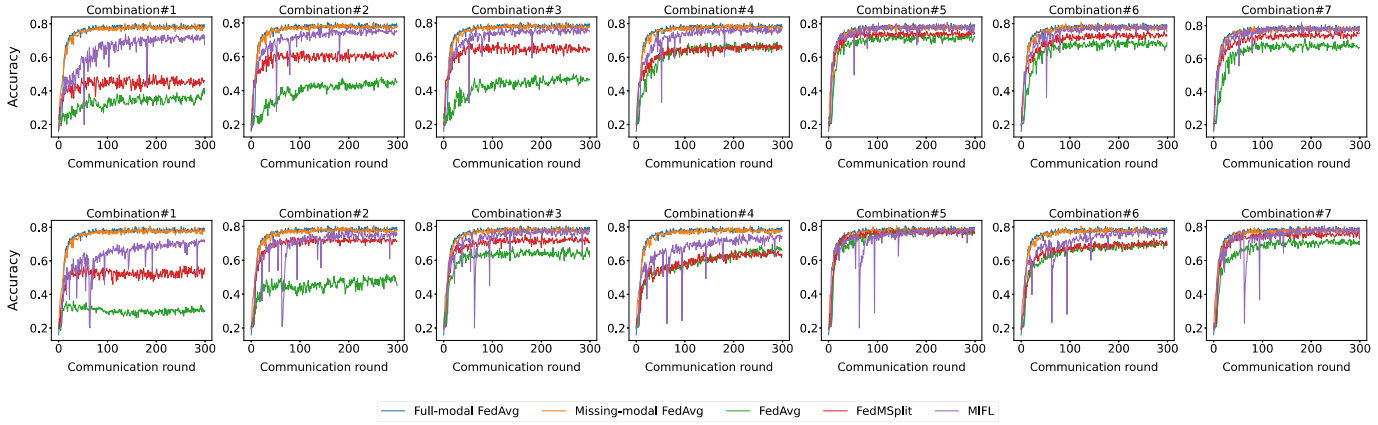

Fig. 5: Performance of different models over the communication round in Experiment 1 and Experiment 2, respectively

TABLE II: Performances of different methods in Experiment 1

Testing combination	Model				
	Full-modal FedAvg	Missing-modal FedAvg	FedAvg	FedMSplit	MIFL
#1	0.807	**0.808**	0.425	0.504	**0.748**
#2		**0.798**	0.488	0.648	**0.780**
#3		**0.801**	0.512	0.686	**0.782**
#4		**0.807**	0.714	0.702	**0.786**
#5		0.802	0.750	0.770	**0.808**
#6		0.807	0.723	0.774	**0.810**
#7		0.799	0.723	0.774	**0.820**
Average	**0.807**	0.803	0.619	0.694	**0.791**

The best results among multimodal-missing techniques are in **bold**. The highest accuracy scores with respect to each testing combination are highlighted in **bold and underlined**.

TABLE III: Performances of different methods in Experiment 2

Testing combination	Model				
	Full-modal FedAvg	Missing-modal FedAvg	FedAvg	FedMSplit	MIFL
#1	0.807	**0.808**	0.367	0.598	**0.758**
#2		**0.798**	0.559	0.750	**0.798**
#3		**0.801**	0.687	0.750	**0.801**
#4		**0.807**	0.707	0.680	**0.790**
#5		0.802	0.787	0.791	**0.804**
#6		**0.807**	0.736	0.734	**0.806**
#7		0.799	0.744	0.782	**0.814**
Average	**0.807**	0.803	0.655	0.726	**0.796**

The best results among multimodal-missing techniques are in **bold**. The highest accuracy scores with respect to each testing modality are highlighted in **bold and underlined**.

Notably, MIFL's performance is comparable to that of the Full-modal FedAvg and Missing-modal FedAvg methods. MIFL outperforms FedAvg trained with 12 modalities in terms of accuracy, with results ranging from 92.69% to 101.6% versus Full-modal and 92.57% to 102.6% versus Missing-modal. The enhancement ranges demonstrate MIFL's effective performance in severe multimodal missing scenarios while exceeding the projected upper bound in less difficult circumstances. Efficient learning of modality-missing information enables MIFL to perform well with fewer modalities, attaining an approximation comparable to the optimal scenario.

2) Ablation Study: The proposed model comprises two main components: Extra Modality Generator, and Contrastive-auxiliary loss function. Section II-B explains the reasoning behind the design of these components.

TABLE IV: Ablation study

Method		**MIFL w/o contrastive loss**	**MIFL w/o Extra Modality Generator**	**MIFL**
Experiment 1 Testing combination	#1	0.499	0.695	**0.748**
	#2	0.641	0.748	**0.780**
	#3	0.701	0.752	**0.782**
	#4	0.662	0.760	**0.786**
	#5	0.776	0.796	**0.808**
	#6	0.757	0.791	**0.810**
	#7	0.760	0.795	**0.820**
	Average	0.685	0.762	**0.791**
Experiment 2 Testing combination	#1	0.550	0.686	**0.758**
	#2	0.700	0.700	**0.798**
	#3	0.716	0.748	**0.801**
	#4	0.656	0.730	**0.790**
	#5	0.776	0.788	**0.804**
	#6	0.712	0.752	**0.806**
	#7	0.732	0.786	**0.814**
	Average	0.692	0.741	**0.796**

In this section, experiments are conducted to empirically support the hypothesis. The details of these experiments are as follows.

- MIFL without Contrastive loss: To assess the impact of the multimodal contrastive-based auxiliary loss, the contrastive loss MultimodalCL($\mathcal{B}$) is excluded from the client's local loss $\mathcal{L}(\mathcal{B})$, while the classification loss $\ell(\mathcal{B})$ and regularization term $R_k(\widehat{w_k})$ are still retained.
- MIFL without Extra Modality Generator: The Extra Modality Generator, which was expected to demonstrate information from absent modalities as well as learning generalizability across different clients in the global model, is eliminated.

Table IV gives information about the accuracy of the aforementioned ablation models. Apparently, the entire proposed framework outperforms the other variants, indicating the positive influence of each architectural decision. In particular, the full version of MIFL surpasses MIFL without contrastive loss by an average percentage of 17.27% for Experiment 1 and 16.02% for Experiment 2. Among all testing combinations,

adding contrastive loss to client training gives an improvement of up to 49.9%, which is higher when evaluating with smaller numbers of modalities, therefore highlighting the advantages of utilizing the contrastive loss in the client's loss. When incorporating the Extra Modality Generator into the global model, there is a relative increase in performance from 1.51% to 14%. Although this block is straightforward, it can nonetheless provide information for completely missing modalities. This endeavor is difficult since embeddings derived from actual samples include valuable and meaningful information. Therefore, this enhancement is considered satisfactory given the complexity of the challenge that this block is addressing. Overall, the architectural proposal and the additional contrastive loss have irreplaceable impacts on the entire MIFL model.

IV. Conclusion

In conclusion, our study introduces a novel Multimodal Federated Learning framework designed to tackle the challenge of heterogeneous clients with missing modalities in their datasets. By leveraging both existing and missing modalities from other clients, our approach enhances local models, promoting generalization across a potentially diverse population of unobserved individuals. This architecture improves the robustness and adaptability of Federated Learning in real-world scenarios, allowing clients to effectively learn from both local and relevant information obtained from other clients, particularly in the presence of missing modalities. Furthermore, the incorporation of a graph-attention-based aggregation method strikes a balance between personalization and collaboration. This technique enables clients' models to retain the advantages of local adaptation while benefiting from global information exchange. Empirical results demonstrate the superior performance of our proposed model in handling missing modalities, providing strong evidence of its efficacy in enhancing Multimodal Federated Learning performance.

Acknowledgement

This research is funded by Hanoi University of Science and Technology (HUST) under grant number T2022-PC-049. This research is also partially supported by NAVER Corporation within the framework of collaboration with the International Research Center for Artificial Intelligence (BKAI), School of Information and Communication Technology, HUST under project NAVER.2022.DA07. The work of BPN was supported in part by the Endeavour Fund – Smart Ideas from the New Zealand Ministry of Business, Innovation and Employment (MBIE) under contract VUW RTVU2301.

References

[1] T. Baltrušaitis et al., "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[2] V. Pérez-Rosas et al., "Utterance-level multimodal sentiment analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 973–982.

[3] A. Zadeh et al., "s," *arXiv preprint arXiv:1707.07250*, 2017.

[4] Z. Jia et al., "Hetemotionnet: two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 1047–1056.

[5] W. Guo et a., "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63 373–63 394, 2019.

[6] C. Zhang et al., "Deep partial multi-view learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2402–2415, 2020.

[7] T. Zhou et al., "Missing data imputation via conditional generator and correlation learning for multimodal brain tumor segmentation," *Pattern Recognition Letters*, vol. 158, pp. 125–132, 2022.

[8] G. Yu et al., "Optimal sparse linear prediction for block-missing multi-modality data without imputation," *Journal of the American Statistical Association*, vol. 115, no. 531, pp. 1406–1419, 2020.

[9] J. Chen and A. Zhang, "Hgmf: heterogeneous graph-based fusion for multimodal data with incompleteness," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2020, pp. 1295–1305.

[10] P. Poklukar et al., "Geometric multimodal contrastive representation learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 782–17 800.

[11] Q. Yu et al., "Multimodal federated learning via contrastive representation ensemble," *arXiv preprint arXiv:2302.08888*, 2023.

[12] B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[13] X. Li et al., "On the convergence of FedAvg on non-IID data," *arXiv preprint arXiv:1907.02189*, 2019.

[14] C. T Dinh et al., "Personalized federated learning with Moreau envelopes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 394–21 405, 2020.

[15] T. Li et al., "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[16] A. Fallah et al., "Personalized federated learning: A meta-learning approach," *arXiv preprint arXiv:2002.07948*, 2020.

[17] Y. J. Cho et al., "Heterogeneous ensemble knowledge transfer for training large models in federated learning," *arXiv preprint arXiv:2204.12703*, 2022.

[18] Q. H. Pham et al., "Sem: A simple yet efficient model-agnostic local training mechanism to tackle data sparsity and scarcity in federated learning," in *2023 Eleventh International Symposium on Computing and Networking (CANDAR)*. IEEE, 2023, pp. 120–126.

[19] N. H. Nguyen et al., "Cadis: Handling cluster-skewed non-iid data in federated learning with clustered aggregation and knowledge distilled regularization," in *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. IEEE, 2023, pp. 249–261.

[20] N. H. Nguyen and P. L. Nguyen et al., "Feddrl: Deep reinforcement learning-based adaptive aggregation for non-iid data in federated learning," in *Proceedings of the 51st International Conference on Parallel Processing*, 2022, pp. 1–11.

[21] F. Liu et al., "Federated learning for vision-and-language grounding problems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 572–11 579.

[22] B. Xiong et al., "A unified framework for multi-modal federated learning," *Neurocomputing*, vol. 480, pp. 110–118, 2022.

[23] Y. Zhao et al., "Multimodal federated learning on IoT data," in *2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI)*. IEEE, 2022, pp. 43–54.

[24] J. Chen and A. Zhang, "FedMSplit: Correlation-adaptive federated multi-task learning across multimodal split networks," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 87–96.

[25] P. Wagner et al., "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, no. 1, p. 154, 2020.

[26] B. McMahan et al., "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017)*, ser. Proceedings of Machine Learning Research, vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282.