Leveraging Foundation Models for Multi-modal Federated Learning with Incomplete Modality

Liwei Che[1,2], Jiaqi Wang[2], Xinyue Liu[3], and Fenglong Ma[2] (✉)

[1] Rutgers University, Piscataway NJ 08854, USA
lw.che@rutgers.edu
[2] Pennsylvania State University, University Park PA 16802, USA
{jawang,fenglong}@psu.edu
[3] Dalian University of Technology, Dalian Liaoning 116621, China
xyliu@dlut.edu.cn

Abstract. Federated learning (FL) has obtained tremendous progress in providing collaborative training solutions for distributed data silos with privacy guarantees. However, few existing works explore a more realistic scenario where the clients hold multiple data modalities. In this paper, we aim to solve a novel challenge in multi-modal federated learning (MFL) – modality missing – the clients may lose part of the modalities in their local data sets. To tackle the problems, we propose a novel multi-modal federated learning method, **Fed**erated **M**ulti-modal contrasti**V**e training with **P**re-trained completion (FedMVP), which integrates the large-scale pre-trained models to enhance the federated training. In the proposed FedMVP framework, each client deploys a large-scale pre-trained model with frozen parameters for modality completion and representation knowledge transfer, enabling efficient and robust local training. On the server side, we utilize generated data to uniformly measure the representation similarity among the uploaded client models and construct a graph perspective to aggregate them according to their importance in the system. We demonstrate that the model achieves superior performance over two real-world image-text classification datasets and is robust to the performance degradation caused by missing modality.

Keywords: Federated Learning · Multi-modal Learning

1 Introduction

Federated learning (FL) has emerged as a promising paradigm for training machine learning models on decentralized data [23,38,35,36,37,1]. In many realistic scenarios, the multi-modal data are collected among distributed data silos and stored in a privacy-sensitive manner, such as the examination and diagnostic records of patients in different hospitals and the multimedia data generated on mobile devices. However, most existing federated learning works focus on single modality scenarios (*e.g.*, image or text) with limited capacity for data with heterogeneous formats and properties. Regarding the fast development of multimedia technology and distributed systems, developing a robust and efficient FL framework for multi-modal machine learning tasks is significant.

To date, several early attempts for multimodal federated learning (MFL) [2] have been proposed [18,48,44,42,6,3,45,46]. Some of these approaches [3,44,45] consider scenarios where the federated system contains both uni-modal and multi-modal clients. However, most of these works assume that all modalities are available to all clients, which is a strong assumption that may not hold in real-world situations. For example, content posted on social media often combines images and text, but users may also publish posts containing only images or text. This modality missing problem poses a substantial challenge as it can severely impact the model's learning ability and performance.

In this paper, we aim to address this general and realistic problem of **modality missing**, where clients share the same modality combinations, but some multi-modal instances lack part of the modality data. For example, a client holds 1000 image-text pairs, while 200 of them only have image data, and 300 instances have only text data. A few existing works [22,20] focus on the modality incompleteness problem. However, they either only consider text missing in the vision-language learning task or deal with sensor signals that are similar in format. We believe that an advanced MFL framework should be robust to modality incomplete training data and maintain satisfactory performance.

To resolve those challenges, we proposed a multi-modal federated learning framework, namely **Fed**erated **M**ulti-modal contrasti**V**e training with **P**re-trained completion (FedMVP), which uses frozen pre-trained models as the teachers to support the learnable multi-modal joint encoder module for efficient multi-modal representation learning and to generate informative synthetic data. To enhance the model resilience to the performance degradation caused by modality missing, we utilize the cross-modal generation ability of the recently proposed pre-trained models [27,15,14] to complete the missing modalities. To further improve the representation learning performance, we proposed an efficient knowledge-transferring method to transfer the representation knowledge from the pre-trained large models to our multi-modal joint learning module. This knowledge-transferring method can alleviate the conflict between the massive data and computing costs requirements for training and fine-tuning of pre-trained large models and the limited resources of federated learning clients. The proposed framework is competent in integrating various pre-trained models with affordable communication costs. As shown in Table 1, compared to the most costly baseline FedViLT, the FedMVP reduces the communication cost by **26.7**$\times$ and computation FLOPS by **15.5**$\times$. The pre-trained foundation models will play as the frozen data encoders to transform the original data into high-quality representations, which play an important role in the contrastive-manner training process for the multi-modal joint encoder module.

We summarize our **contributions** as follows: (1) We proposed a novel MFL framework that integrates pre-trained large-scale models to conduct efficient multi-modal representation learning and is robust to the modality missing challenge. Our proposed method shows superior performance on two multi-modal classification benchmarks under both IID and non-IID settings. (2) To efficiently transfer the learnable representation knowledge from the pre-trained model to

Table 1. Comparison between FedMVP and baselines in terms of #FLOPS (Floating Point Operations Per Second) and #transmitted parameters per round.

Method	#FLOPS	#Parameters
FedViT [8]	$22.6G$	$86.4M$
FedBERT [7]	$38.1G$	$110.1M$
FedCLIP [27]	$60.7G$	$197.2M$
FedViLT [21]	$55.9G$	$298.6M$
MMFed [42]	$1.4G$	$4.49M$
FedMVP	$3.6G$	$11.2M$

the multi-modal joint module under the resource-limited scenario, we proposed a Multi-modal Contrastive Matching (MCM) loss and a Representation Aligned Margin (RAM) loss, which effectively improve the model performance with severe modality missing up to 80%. (3) Instead of aggregating the models based on the data distribution or the model architecture, we propose a novel aggregation algorithm for the MFL server aggregation based on the representation abilities among the client models.

2 Related Work

Multi-modal Federated Learning (MFL). MFL is still in its early stages of development. Some of the most existing works [42,48,18] focus on exploring task-specific approaches with complete modalities. In [42], the authors propose a multi-modal federated learning framework for multi-modal activity recognition with a local co-attention module to fuse multi-modal features. [5] gives a detailed analysis of the convergence problem of MFL with late fusion methods under the Non-IID setting. [45,3,44] adapt modality-wise encoders to tackle the MFL system with both uni-modal and multi-modal clients. However, few of them explore the scenario where multi-modal data are incomplete, which may cause significant performance degradation.

Modality Missing in Multi-modal Learning. As a widely existing challenge in the realistic scenario, handling modality missing has drawn the attention of the multi-modal learning community. Some early works [25,39,30] build their methods based on conditional VAE to capture the multi-modal distribution for the cross-modal generation. [33] as one of the recent works utilizes cross-modal fusion to improve the model robustness for modality missing in testing. [29] proposes a contrastive framework for learning both paired and unpaired data. In [22], the authors leverage Bayesian meta-learning to reconstruct pseudo text input from image input to resolve the missing modality issue. Instead of training a generative model from scratch, we utilize the large-scaled pre-trained model [14,15] and prompt augmentation to achieve effective cross-modal generation for completing the missing data pairs.

Vision-language Pre-training. Represented by CLIP [27] and ALIGN [12], the large-scale Vision and Language Pre-training (VLP) models have demon-

strated their surprising performance in many downstream vision-language learning tasks[10] and strong adaptability to new scenarios. A few works have taken the first steps towards incorporating federated learning with pre-training techniques. In [32], the authors propose a splitting learning-based framework for training large-scale models like BERT in federated learning systems. PromptFL [9] allows the clients to train shared soft prompts collaboratively using CLIP [27] to provide strong adaptation capability to distributed users tasks. [4,19,41,40] are trying to explore the efficient methods for lightweight and fast adaptation of pre-trained models. [31] proposes FedPCL to transfer shared knowledge among the clients based on prototype contrastive learning. In this work, instead of fine-tuning the large-scale pre-trained models or splitting the model into multiple modules, we conduct effective knowledge transferring to enhance the representation learning performance of a lightweight local module.

Multi-modal Contrastive Learning. Contrastive learning is widely used in the self-supervised learning field, where the learned representations will be assigned to positive and negative samples based on the class belongings. As for its application in multi-modal learning [16,17,47], instead of using spatial or temporal transforming to a single instance, the positive pairs are defined as the samples with the same ID or time window. In [47], the authors propose CrossCLR to improve the quality of learned joint embedding from multi-modal data with a novel contrastive loss, which utilizes both inter-modality and intra-modality alignment. [26] extends the multi-modal contrastive learning to efficiently align the cross-modal representations. Inspired by the predecessors, we adopt a multi-modal contrastive loss to improve the quality of the learned multi-modal joint representations based on the modality-specific representation encoded by the frozen pre-trained models.

3 Methodology

To explore multi-modal data in federated systems, we propose **FedMVP** for MFL with the robustness of modality missing during training. As illustrated in Figure 1, the proposed FedMVP contains four main modules for effective MFL, including *Modality Completion Module*, *Multi-modal Joint Learning Module*, *Knowledge Transferring via Contrastive Training*, *CKA-based Aggregation*.

3.1 Problem Formulation

Multi-modal Federated Learning. In an MFL system, there exist N clients aiming to collaboratively train a global model w_G for multi-modal representation learning. For client n, its local data set $\mathcal{D}_n = \{(X_i, y_i)\}_{i=1}^{|\mathcal{D}_n|}$ contains $|\mathcal{D}_n|$ image-text pairs denoted as $X_i = \{x_i^I, x_i^T\}$, i.e., the i-th image data x_i^I and text data x_i^T. y_i is the corresponding label. A data instance is denoted as $X_i = \{x_i^I\}$ or $X_i = \{x_i^T\}$ if modality missing happens. Each local model w_n performs on the local task $F_n(\cdot; w_n) : \mathbb{R}^n \to \mathbb{R}^d$ and collaborates with other clients for the

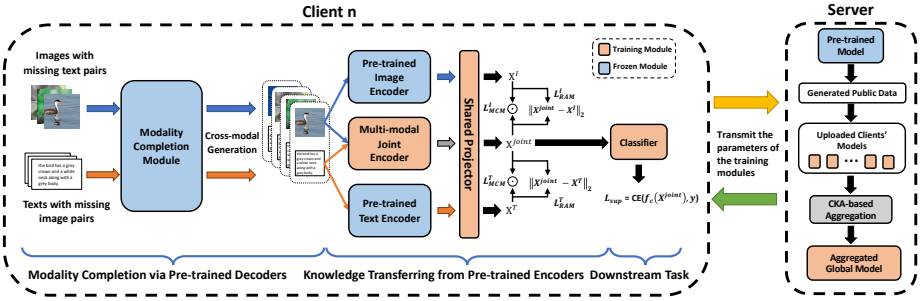

Fig. 1. The overview of the proposed FedMVP framework.

global task $F_G(\cdot; w_G) : \mathbb{R}^{d_G} \to \mathbb{R}^d$. Formally, the global objective of MFL for the image-text classification problem is defined as

$$\min \; L_G(F_G(\cdot; w_G)) = \min \; \sum_{n=1}^{N} \gamma_n L_n(F_n(\mathcal{D}_n; w_n)) \tag{1}$$

where γ_n is the aggregation weights, and L_n is the local loss function.

3.2 Local Data Preprocessing

A pre-trained foundation model is deployed on both the server side and client side, which consists of an image encoder $f_E^I(\cdot)$ and a text encoder $f_E^T(\cdot)$ for representation extraction, an image decoder $f_D^I(\cdot)$ and a text decoder $f_D^T(\cdot)$ for the cross-modal generation. Notably, all the parameters of the pre-trained models are frozen and will not be transmitted between the server and clients. We will explain the pre-trained model we used below, as well as the details of the local training process.

Modality Completion Module. To solve the performance drop problem caused by modality missing, the modality completion module utilizes the cross-modal generation ability of the pre-trained model to complete the missing part of multi-modal data. We use DALLE2 [28] for text-to-image generation, and BLIP2 [14] for image-to-text generation. Inspired by [27], we use designed prompts to improve the generation quality of the modality completion module.

Prompt Augmented Text-to-image Generation. Given an image-text pair X_i with only text data x_i^T, the modality completion module could generate an image from a text prompt. To avoid the semantic ambiguities caused by synonyms and polysemy in the text data and label name. Instead of directly using text data x_i^T as the input, we adopt a coarse-to-fine prompt to augment the generation. The prompt template is *"A photo of {fine-grained label}, a kind of {class label}, {text description}"*, which helps the pre-trained models to better understand the characteristics of the generation target and improve the semantic correlation between the text prompt and generated image. Figures 2 and 3 show examples

(a) original text (b) label only (c) prompt (d) original image

Fig. 2. Examples of the generated "snapdragon" images.

(a) original text (b) label only (c) prompt (d) original image

Fig. 3. Examples of the generated "yellowthroat" images.

with different inputs to generate the classes "snapdragon" and "yellow throat" on the Oxford Flower and CUB-200 datasets, where our designed prompt gives high-quality fake images that are close to the original ones.

Accordingly, we obtain the image generation prompt based on the original text data, and the process is denoted as $p^T(x_i^T)$. The augmented prompt $p^T(x_i^T)$ will firstly be decomposed by text encoder $f_E^T(\cdot)$, then passed to image decoder f_D^I for generating the synthetic image $\hat{x}_i^I$, i.e.$\hat{x}_i^I = f_D^I(f_E^T(p^T(x_i^T)))$.

Prompt Augmented Image-to-text Generation. For the image-to-text generation, considering the original text data contains detailed descriptions of the image pair, the direct image captioning result may not be able to cover the fine-grained text details. Therefore, we adapt both the visual question answering (VQA) and image captioning functions of the pre-trained model to generate text pairs $\hat{x}_i^T$ for the image input. Specifically, with a given image input x_i^I, the modality completion module first performs the VQA task over three serial question prompts to get fine-grained descriptions of the image. For instance, given prompt input *"What is the color of the petals?"* for a flower image with the pink pedal, the response answer could be *"Pink"*. After obtaining the answers to the three question prompts, we combine them with the image captioning outcome as the final synthetic text, e.g., *"A photo of {flower}, with {pink} petals and {white} pistils,{there is a pink flower with a yellow center in the middle of the picture}"*. We show examples of image-to-text generation in Table 2.

To better understand the model design and avoid notation confusion, we use completed image-text pair $X_i = \{x_i^I, x_i^T\}$ in the following sections to illustrate how data is processed in FedMVP.

Table 2. Image-to-text completion examples from CUB-200 and Oxford Flower.

Type	Text	Original Image
Original Text	this flower is yellow in color, and has petals that are layered.	
Synthetic Text	A flower with yellow petals and yellow pistil. yellow flower with water droplets on it in a garden	
Original Text	this bird is yellow and black in color, and has a stubby black beak.	
Synthetic Text	A bird with black wings and yellow belly. yellow and black bird perched on a cattails plant in a marsh	

Modality-specific Representations. The foundation models are believed to have extraordinary representation extraction ability since they are trained with millions of data instances. Thus, we obtain the image-specific embedding and text-specific embedding via the pre-trained encoders. Specifically, we use the pre-trained Vision Transformer(ViT) [8] with the patch size of 16×16 as the image-specific encoder to generate high-quality embedding from image input. The image-specific embedding $\mathbf{X}^I$ is encoded via the pre-trained image encoder $f_E^I(\cdot)$ and then mapping to the multi-modal latent space via a shared projection head $f_{shared}(\cdot)$, i.e., $\mathbf{X}^I = f_{shared}(f_E^I(\mathbf{x}^I)) \in \mathbb{R}^{d_{latent}}$. Similarly, we get the text-specific embedding $\mathbf{X}^T$ from the pre-trained BERT model [7], where $\mathbf{X}^T = f_{shared}(f_E^T(\mathbf{x}^T)) \in \mathbb{R}^{d_{latent}}$.

3.3 Local Training

Multi-modal Joint Learning Module. The multi-modal joint learning module contains a joint encoder $f_E^{Joint}(\cdot)$ designed to efficiently fuse the image-text information into a complete view. It consists of a cross-modal fusion layer and follows attention-based embedding layers.

Pre-processing. Given an image-text pair $\{\mathbf{x}^I, \mathbf{x}^T\}$ as input, we use a non-overlapped patch embedding layer and the pre-trained text encoder $f_E^T(\cdot)$ to get the patch sequence $\mathbf{I}_{com}$ and text embedding $\mathbf{T}_{com}$, both belongs to the common dimension d_{com}.

Cross-modal Fusion. After the positional embedding operation, both the image and text embeddings are fed into the cross-modal fusion layer, which contains a vision-to-language attention module and a language-to-vision attention module. Both modules are based on the cross-modal attention [33], which can effectively fuse the representation between the two input modality embeddings.

We take the image-to-text embedding $\mathbf{X}^{I\to T}$ to show the cross-modal attention:

$$\mathbf{X}^{I\to T} = CM_{I\to T}(\mathbf{I}_{com}, \mathbf{T}_{com}) = softmax(\frac{W_{Q_I}\mathbf{I}_{com}W_{K_T}^{\mathsf{T}}\mathbf{T}_{com}^{\mathsf{T}}}{\sqrt{d_{com}}})W_{V_T}. \quad (2)$$

Similarly, we can get text-to-image embedding $\mathbf{X}^{T\to I}$. The obtained $\mathbf{X}^{I\to T}$ and $\mathbf{X}^{T\to I}$ will be concatenated together and projected to the latent space as the final joint embedding via the shared projection head $f_{shared}(\cdot)$ and a self-attention layer as follows:

$$\mathbf{X}^{joint} = f_{shared}(SelfAttention(\mathbf{X}^{I\to T} \oplus \mathbf{X}^{T\to I})). \quad (3)$$

We now obtain the image-specific embedding $\mathbf{X}^I$, text-specific embedding $\mathbf{X}^T$, and joint embedding $\mathbf{X}^{joint}$ in the same latent space $\mathbb{R}^{d_{latent}}$.

Knowledge Transferring from Pre-trained Model. The training data of large-scale models in the pre-training stage is neither available nor affordable for distributed silos to process, making the fine-tuning and traditional knowledge distillation [11] of large-scale models impractical under the MFL scenario. In order to transfer the rich representation knowledge from the pre-trained model, we propose *Multi-modal Contrastive Matching (MCM) Loss* and *Representation Aligned Marginal (RAM) Loss* to improve the representation learning performance of the joint encoding module.

Multi-modal Contrastive Matching Loss. To obtain a high-quality joint representation, we utilize the idea of contrastive learning to closer the joint embedding with its corresponding modality-specific embedding and distance it from the embedding of the other categories in the latent space. Let $s_c(x_i, x_j)$ represent the cosine similarity between two embedding, x_i and x_j, and $\tau \in (0, 1]$ be the temperature hyperparameter. The corresponding scaled similarity is defined as: $sim(x_i, x_j) = \exp(\frac{s_c(x_i,x_j)}{\tau})$.

Given a batch of embedding $\mathcal{B} = \{\mathbf{X}_i^T, \mathbf{X}_i^I, \mathbf{X}_i^{joint}\}_{i=1}^{|\mathcal{B}|}$, the positive pair for the contrastive learning is defined as the joint embedding with its corresponding modality-specific embedding, i.e., $(\mathbf{X}_i^T, \mathbf{X}_i^{joint})$ and $(\mathbf{X}_i^I, \mathbf{X}_i^{joint})$. The other ways of pairing will be treated as negative pairs, denoted as:

$$\Omega_i^m = \sum_{i\neq j}(sim(\mathbf{X}_i^M, \mathbf{X}_j^M) + sim(\mathbf{X}_i^M, \mathbf{X}_j^{joint}) + sim(\mathbf{X}_i^{joint}, \mathbf{X}_j^{joint})), \quad (4)$$

where $M \in \{I, T\}$ indicates the modality type. We define the multi-modal contrastive matching (MCM) loss of all data embedding as follows:

$$L_{MCM}(\mathcal{B}) = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\log\left(\frac{sim(\mathbf{X}_i^T, \mathbf{X}_i^{joint})}{\Omega_i^T} + \frac{sim(\mathbf{X}_i^I, \mathbf{X}_i^{joint})}{\Omega_i^I}\right). \quad (5)$$

Representation Aligned Margin Loss. We propose the *Representation Aligned Margin* (RAM) loss to further enrich the joint representation via pre-trained knowledge to close the semantic gap between the joint embedding and the

modality-specific embeddings. We use the classification loss derived from the embeddings to evaluate its representation quality. For the i-th data sample, the supervised classification loss of one of its corresponding embeddings is denoted as $L_{sup}^{M}(i) = CE(f_c(\mathbf{X}_i^M), y_i)$.

Intuitively, embeddings with lower cross-entropy losses contain more informative features from the raw data. With an embedding batch $\mathcal{B}$, the RAM loss aligns joint embedding with image and text embedding separately, if the modality-specific embedding has better representation. Thus, the RAM loss is defined as:

$$L_{RAM}(\mathcal{B}) = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(L_{RAM}^{I}(i) + L_{RAM}^{T}(i)\right), \tag{6}$$

$$L_{RAM}^{M}(I) = \begin{cases} \|\mathbf{X}_i^{joint} - \mathbf{X}_i^{M}\|_2, & \text{if } L_{sup}^{M}(i) < L_{sup}^{joint}(i) \\ 0, & \text{otherwise} \end{cases}, \tag{7}$$

where $\mathbf{X}_i^M$ and $\mathbf{X}_i^{joint}$ are all derived from the i-th sample in the batch, and $|\mathcal{B}|$ is the batch size. The L2 norm is denoted by $\|\cdot\|_2$.

Classification Loss. A two-layer linear classifier $f_C(\cdot)$ will serve as the classifier using only joint embedding as input. The supervised classification loss L_{sup} of client n can be obtained:

$$L_{sup}(\mathcal{B}) = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|} CE\left(f_C\left(\mathbf{X}_i^{joint}; \boldsymbol{\omega}_n\right), y_i\right), \tag{8}$$

where $f_C(\cdot)$ denotes the classifier model of client n, $CE(\cdot)$ is the cross-entropy loss function, and y_i is the corresponding label of i-th joint embedding $\mathbf{X}_i^{joint}$.

Total Loss. The final local training loss of client k in FedMVP is:

$$L_{local}(\mathcal{D}_k) = L_{sup}(\mathcal{D}_k) + L_{MCM}(\mathcal{D}_k) + L_{RAM}(\mathcal{D}_k), \tag{9}$$

At each communication round, each client will upload the parameters of the multi-modal joint learning module and classifier to the server for further global aggregation.

3.4 Server Aggregation

Previous works tend to aggregate based on the modality type held by the clients [3,43], share public dataset [44], or model structure [45], which may lead to data privacy leakage and lacking uniformity. To better enhance the representational ability of the global model, we propose a server aggregation method based on the similarity of model output representations.

At the beginning of the aggregation phase, the server-side pre-trained model will automatically generate m synthetic data pairs X_m, where the data amount m is equal to the number of classes of the dataset. Given an uploaded client model, its output representations with generated data are defined as:

$$\mathbf{X}_\omega = [F_\omega(X_1), \ldots, F_\omega(X_m)]^T \in \mathbb{R}^{m \times d_{out}}. \tag{10}$$

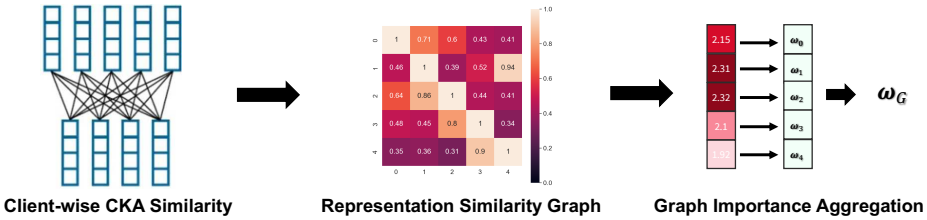

Fig. 4. CKA-based Server Aggregation

To measure the similarity of the model representations among the clients, we utilize the centered kernel alignment (CKA) metric [13] based on the output representations from upload models, which is defined as follows:

$$s_{ij}(\omega_i, \omega_j) = \frac{Cov(\mathbf{X}_{\omega_i}, \mathbf{X}_{\omega_j})}{\sqrt{Cov(\mathbf{X}_{\omega_i}, \mathbf{X}_{\omega_i}) Cov(\mathbf{X}_{\omega_j}, \mathbf{X}_{\omega_j})}}, \tag{11}$$

where $Cov(X, Y) = (m-1)^2 tr(XX^T H_m Y Y^T H_m)$, H_m is the centering matrix, $tr(\cdot)$ denotes the matrix trace, m represents the number of input represents.

With the calculated representation similarity scores, the server constructs a representation similarity graph to illustrate the relationship among clients, as shown in Figure 4. The importance of each client in the representation similarity graph is determined by the sum of its similarity score with all the other clients.

$$\gamma_i^t = softmax([s_1, \ldots, s_i, \ldots, s_K]), \tag{12}$$

where K is the number of clients who participate in the t-th aggregation, $s_i = \sum_{j=1}^{K-1} s_{ij}$ is the collection of the graph importance of all K clients. Finally, the global model is weighted and aggregated based on the clients' graph importance γ_i^t as follows:

$$w_G^t = \sum_{i=1}^{K} \gamma_i^t w_i^t. \tag{13}$$

4 Experiments

4.1 Experiment Setting

Datasets. We evaluate the proposed FedMVP on two multi-modal fine-grained categorization datasets, *The Caltech-UCSD Birds-200-2011 (CUB-200) dataset* [34] and *Oxford Flower*[24]. Both contain paired image-text data, and each image has 10 related descriptive text. CUB-200 has 200 bird classes with 10610 training image-text instances and 1178 for testing. Oxford Flower has 102 flower classes, a training size of 7370, and a testing size of 819.

Data Distribution Setting. For **Independent Identically Distribution (IID)** setting, we equally distribute the training data to 10 clients with

Table 3. Evaluating the impact of incomplete modality on CUB-200 and Oxford Flower datasets under IID setting. β indicates the missing ratio of the training set.

Methods	CUB-200			Oxford Flower		
	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.8$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.8$
FedViT	74.71%	67.12%	60.33%	92.15%	84.52%	76.64%
FedBERT	66.76%	58.98%	52.54%	74.23%	70.72%	67.81%
FedCLIP	75.73%	69.68%	63.41%	91.12%	86.32%	78.55%
FedViLT	76.29%	70.28%	64.11%	92.67%	88.31%	81.52%
MMFed	63.15%	57.48%	51.60%	72.91%	69.43%	64.05%
FedMVP(Ours)	**77.89%**	**74.46%**	**70.31%**	**93.19%**	**91.28%**	**89.32%**

random selection. Each client will hold the same quantity of local data with a balanced category distribution. To simulate the **non-IID** scenario in federated systems, we divide the training data set into C shards according to the data set categories, i.e., 200 shards for CUB-200-2011 dataset and 102 shards for the Oxford Flower dataset. With fixed 10 clients, the data shards are randomly and equally distributed to clients.

Modality Missing Setting. We set $\beta \in [0, 1]$ as the missing ratio. For example, given a constant $\beta = 0.3$, 30% randomly selected image-text pairs will lose either image or text data in equal chances. We select $\beta = 0.3, 0.5, 0.8$ to conduct our experiments, and the number of missing images and texts is the same.

Training Setting. With fixed 10 clients, the total communication round is 200. In each communication round, the clients will perform 10 epochs for local training with their own local datasets and the server will randomly select 70% of clients for aggregation. We choose AdamW as the optimization function with a scheduler-controlled learning rate $2e - 5$. We adopt the warm-up scheduler and cosine annealing scheduler for the training process as well.

Baselines. Since the existing approaches for addressing modality missing in multi-modal federated learning are relatively limited, we choose **FedViT**, **FedBERT** as the uni-modal baseline and **FedCLIP**, **FedViLT**, **MMFed** as the multi-modal baseline. FedViT [8], FedBERT [7], FedCLIP [27] and FedViLT [21] are using large-scale foundation models pre-trained with millions of data as the local models. These large models are fine-tuned on the local data and upload all the parameters to the server for aggregation. MMFed [42] is a federated multi-modal learning method without leveraging foundation models. FedViLT [21] is designed specifically for modality missing. *Please refer to Appendix for details of the implementation.*

4.2 Empirical Results

Results of the IID Setting. Table 3 shows the superior performance of FedMVP across different missing ratios under the IID setting on both CUB-200 and

Table 4. Evaluating the impact of incomplete modality on CUB-200 and Oxford Flower datasets under the non-IID setting. β indicates the missing ratio of the training set.

Methods	CUB-200			Oxford Flower		
	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.8$	$\beta = 0.3$	$\beta = 0.5$	$\beta = 0.8$
FedViT	67.05%	61.17%	50.39%	86.25%	78.30%	70.03%
FedBERT	59.31%	51.14%	43.67%	68.43%	62.01%	57.16%
FedCLIP	67.63%	61.72%	56.78%	85.01%	80.13%	72.91%
FedViLT	69.19%	65.26%	58.34%	86.96%	81.63%	73.32%
MMFed	57.55%	51.12%	42.14%	65.90%	59.26%	52.79%
FedMVP(Ours)	**72.62%**	**69.73%**	**66.44%**	**88.54%**	**84.78%**	**82.47%**

Table 5. Evaluating the robustness of the methods over different test sets. *image only* and *text only* indicate the test set only contains either image or text. All the methods are trained over train set WITHOUT modality missing.

Methods	CUB-200			Oxford Flower		
	image only	text only	complete	image only	text only	complete
FedCLIP	56.47%	47.30%	79.73%	64.11%	53.59%	94.12%
FedViLT	64.55%	52.08%	**82.29%**	76.71%	60.91%	**96.67%**
MMFed	7.94%	13.07%	65.28%	26.37%	40.90%	74.89%
FedMVP(Ours)	**70.39%**	**64.44%**	80.79%	**80.82%**	**73.50%**	94.27%

Oxford Flower datasets. Observably, all models exhibit a decline in accuracy with an increase in the missing ratio (β). FedMVP outperforms baseline methods consistently and demonstrates exceptional resilience to performance degradation due to missing modalities. For instance, on the CUB-200 dataset, FedMVP's accuracy margin over the next best-performing model, FedViLT, widens from about 1.6% at $\beta = 0.3$ to 6.2% at $\beta = 0.8$. A similar trend is observed on the Oxford Flower dataset, with the margin increasing from 0.52% to 7.8%. The rate of performance degradation of FedMVP is notably slower than the other models. Specifically, as β increases from 0.3 to 0.8, the accuracy of FedMVP drops by merely 7.58% and 3.87% on the CUB-200 and Oxford Flower datasets, respectively. In contrast, FedViT witnesses larger drops of 14.38% and 15.51%.

Results of the Non-IID Setting. The non-IID experimental results, presented in Table 4, all methods experience a significant decrease in accuracy compared to the IID setting, including FedMVP. The proposed FedMVP consistently outperforms the other methods across the settings. FedMVP has minimal performance degradation caused by non-IID compared to the baseline methods, with no more than 5% drop on CUB-200 and no more than 7% on Oxford Flower. Despite the increasing missing ratio from $\beta = 0.3$ to $\beta = 0.8$, FedMVP maintains a substantial lead in accuracy on both datasets. For instance, even with $\beta = 0.8$, FedMVP achieves an accuracy of 66.44% and 82.47% on the

Table 6. Ablation study on both CUB-200 and Oxford Flower datasets with $\beta = 0.3$ under non-IID setting; wo/MCM denoting MCM Loss excluded; wo/RAM excludes RAM loss; wo/Completion refers to training without modality completion module; wo/CKA indicates server aggregation as FedAvg.

Model	CUB-200	Oxford Flower
FedMVP	72.62%	88.54%
-wo/MCM	66.87%	81.44%
-wo/RAM	68.25%	83.60%
-wo/Completion	67.49%	81.87%
-wo/CKA	70.11%	85.01%

CUB-200 and Oxford Flower datasets, respectively, confirming its robustness to modality incompleteness under non-IID settings. Notably, the performance margin between FedMVP and baseline is further widened compared to the IID setting. For instance, on the Oxford Flower dataset, as $\beta = 0.8$, the accuracy of FedMVP is 29.68% higher than MMFed compared to 25.27% under IID.

Results of Single-modality Testing. Shown in Table 5, all methods experience significant performance drops when tested with only one modality (image or text). FedMVP shows the best resilience, achieving the highest accuracy in both image-only and text-only scenarios across datasets. FedViLT[21] performs best with complete data since it has 26.7× more parameters than FedMVP and is pre-trained over millions of pre-training data. It holds second place in single-modality tests. FedCLIP's performance is limited by local dataset size but benefits from separate ViT and BERT encodings. MMFed suffers the most due to its co-attention mechanism and performs better in text-only testing due to its integrated BERT. In summary, FedMVP demonstrates robustness in both training and testing under missing modalities.

Ablation Study. The results in Table 6 show that all the modules in the FedMVP model significantly contribute to its performance. Experimental results show that MCM loss and RAM loss can effectively improve the quality of the representation generated by the multi-modal joint encoder and enhance the final performance of the model by transferring pre-trained knowledge through representation learning. The modality completion module can supplement the data by providing additional training information using the transferable knowledge of the pre-trained model. Furthermore, the experimental results suggest that CKA similarity can effectively measure the importance of the representation learned by each client's local model and can improve aggregation performance compared to traditional average aggregation.

5 Conclusion

In conclusion, we proposed the FedMVP framework to tackle modality missing, a widely existing real-world challenge, where part of the multi-modal data is incom-

plete and unaligned. Our framework utilizes large-scale pre-trained models with frozen parameters for modality completion and representation knowledge transfer at each client. It provides a solution for integrating large-scale pre-trained models to empower the federated system with robustness towards modality incompleteness. The experiments on the real-world image-text pair benchmark demonstrated the effectiveness of our proposed method. The proposed FedMVP framework shows great potential in addressing the missing modality and unified representation learning challenges of multi-modal federated learning. We hope this work can provide inspiration for future research in this field.

Acknowledgments. This study was partially supported by the National Science Foundation under Grant No. 2238275, 2333790, and 2348541.

References

1. Che, L., Long, Z., Wang, J., Wang, Y., Xiao, H., Ma, F.: Fedtrinet: A pseudo labeling method with three players for federated semi-supervised learning. In: 2021 IEEE Big Data. pp. 715–724 (2021)
2. Che, L., Wang, J., Zhou, Y., Ma, F.: Multimodal federated learning: A survey. Sensors **23**(15) (2023)
3. Chen, J., Zhang, A.: Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In: ACM SIGKDD. p. 87–96 (2022)
4. Chen, J., Xu, W., Guo, S., Wang, J., Zhang, J., Wang, H.: Fedtune: A deep dive into efficient federated fine-tuning with pre-trained transformers (2022)
5. Chen, S., Li, B.: Towards optimal multi-modal federated learning on non-iid data with hierarchical gradient blending. In: IEEE INFOCOM 2022-IEEE Conference on Computer Communications. pp. 1469–1478. IEEE (2022)
6. Cobbinah, B.M., Sorg, C., Yang, Q., Ternblom, A., Zheng, C., Han, W., Che, L., Shao, J.: Reducing variations in multi-center alzheimer's disease classification with convolutional adversarial autoencoder. Medical image analysis **82**, 102585 (2022)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Guo, T., Guo, S., Wang, J., Xu, W.: Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. arXiv preprint arXiv:2208.11625 (2022)
10. He, X., Peng, Y.: Fine-grained visual-textual representation learning. IEEE Transactions on Circuits and Systems for Video Technology **30**(2), 520–531 (2019)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015)
12. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)

13. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International Conference on Machine Learning. pp. 3519–3529. PMLR (2019)
14. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
15. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
16. Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., Wang, H.: Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. arXiv preprint arXiv:2012.15409 (2020)
17. Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. arXiv preprint arXiv:2203.02053 (2022)
18. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Federated learning for vision-and-language grounding problems. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11572–11579 (2020)
19. Lu, W., Hu, X., Wang, J., Xie, X.: Fedclip: Fast generalization and personalization for clip in federated learning. arXiv preprint arXiv:2302.13485 (2023)
20. Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X.: Are multimodal transformers robust to missing modality? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18177–18186 (June 2022)
21. Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X.: Are multimodal transformers robust to missing modality? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18177–18186 (2022)
22. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: Smil: Multimodal learning with severely missing modality. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2302–2310 (2021)
23. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
24. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics and Image Processing (Dec 2008)
25. Pandey, G., Dukkipati, A.: Variational methods for conditional multimodal deep learning. In: 2017 international joint conference on neural networks (IJCNN). pp. 308–315. IEEE (2017)
26. Poklukar, P., Vasco, M., Yin, H., Melo, F.S., Paiva, A., Kragic, D.: Geometric multimodal contrastive representation learning. In: International Conference on Machine Learning. pp. 17782–17800. PMLR (2022)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
28. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv:2204.06125 (2022)
29. Shi, Y., Paige, B., Torr, P.H., Siddharth, N.: Relating by contrasting: A data-efficient framework for multimodal generative models. arXiv preprint arXiv:2007.01179 (2020)

30. Suzuki, M., Nakayama, K., Matsuo, Y.: Joint multimodal learning with deep generative models. arXiv preprint arXiv:1611.01891 (2016)
31. Tan, Y., Long, G., Ma, J., Liu, L., Zhou, T., Jiang, J.: Federated learning from pre-trained models: A contrastive learning approach. arXiv:2209.10083 (2022)
32. Tian, Y., Wan, Y., Lyu, L., Yao, D., Jin, H., Sun, L.: Fedbert: When federated learning meets pre-training. ACM Trans. Intell. Syst. Technol. **13**(4) (2022)
33. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the conference. Association for Computational Linguistics. Meeting. vol. 2019, p. 6558. NIH Public Access (2019)
34. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: Caltech-ucsd birds-200-2011 (cub-200-2011). Tech. rep. (2011)
35. Wang, J., Chen, Y., Wu, Y., Das, M., Yang, H., Ma, F.: Rethinking personalized federated learning with clustering-based dynamic graph propagation. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 155–167 (2024)
36. Wang, J., Qian, C., Cui, S., Glass, L., Ma, F.: Towards federated covid-19 vaccine side effect prediction. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 437–452. Springer (2022)
37. Wang, J., Yang, X., Cui, S., Che, L., Lyu, L., Xu, D.D., Ma, F.: Towards personalized federated learning via heterogeneous model reassembly. Advances in Neural Information Processing Systems **36** (2024)
38. Wang, J., Zeng, S., Long, Z., Wang, Y., Xiao, H., Ma, F.: Knowledge-enhanced semi-supervised federated learning for aggregating heterogeneous lightweight clients in iot. In: Proceedings of the 2023 SIAM International Conference on Data Mining (SDM). pp. 496–504. SIAM (2023)
39. Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. Advances in neural information processing systems **31** (2018)
40. Wu, X., Huang, F., Hu, Z., Huang, H.: Faster adaptive federated learning. Proceedings of the AAAI Conference on Artificial Intelligence **37**(9), 10379–10387 (2023)
41. Wu, X., Lin, W.Y., Willmott, D., Condessa, F., Huang, Y., Li, Z., Ganesh, M.R.: Leveraging foundation models to improve lightweight clients in federated learning (2023)
42. Xiong, B., Yang, X., Qi, F., Xu, C.: A unified framework for multi-modal federated learning. Neurocomputing **480**, 110–118 (2022)
43. Yang, X., Xiong, B., Huang, Y., Xu, C.: Cross-modal federated human activity recognition via modality-agnostic and modality-specific representation learning (2022)
44. Yu, Q., Liu, Y., Wang, Y., Xu, K., Liu, J.: Multimodal federated learning via contrastive representation ensemble. In: ICLR (2023)
45. Zhao, Y., Barnaghi, P., Haddadi, H.: Multimodal federated learning on iot data. In: 2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI). pp. 43–54. IEEE (2022)
46. Zhou, Y., Wu, J., Wang, H., He, J.: Adversarial robustness through bias variance decomposition: A new perspective for federated learning. In: CIKM. pp. 2753–2762. ACM (2022)
47. Zolfaghari, M., Zhu, Y., Gehler, P., Brox, T.: Crossclr: Cross-modal contrastive learning for multi-modal video representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1450–1459 (2021)
48. Zong, L., Xie, Q., Zhou, J., Wu, P., Zhang, X., Xu, B.: Fedcmr: Federated cross-modal retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1672–1676 (2021)