Review

Multimodal Federated Learning: A Survey

Liwei Che , Jiaqi Wang , Yao Zhou , Fenglong Ma [*]

Posted Date: 20 July 2023

doi: 10.20944/preprints202307.1420.v1

Keywords: federated learning; multimodal learning; artificial intelligence of things

Article

Multimodal Federated Learning: A Survey

Liwei Che [1], Jiaqi Wang [1], Yao Zhou [2] and Fenglong Ma [1,*]

[1] College of Information Sciences and Technology, Pennsylvania State University, USA
[2] Instacart, USA
[*] Correspondence: fenglong@psu.edu

Abstract: Federated learning (FL) has become a burgeoning and attractive research area, which provides a collaborative training scheme for distributed data sources with privacy concerns. Most existing FL studies focus on taking unimodal data, such as images and text, as the model input and resolving the heterogeneity challenge, i.e., the non-identically distributed (non-IID) challenge caused by data distribution imbalance related to data labels and data amount. In real-world applications, data are usually described by multiple modalities. However, to the best of our knowledge, only a handful of studies have been proposed to improve the system performance by utilizing multimodal data. In this survey paper, we identify the significance of this emerging research topic – multimodal federated learning (MFL) and perform a literature review on the state-of-art MFL methods. Furthermore, we categorize multi-modal federated learning into congruent and incongruent multimodal federated learning based on whether all clients possess the same modal combinations. We investigate the feasible application tasks and related benchmarks for MFL. Lastly, we summarize the promising directions and fundamental challenges in this field for future research.

Keywords: federated learning; multimodal learning; artificial intelligence of things

1. Introduction

In many realistic scenarios, data are usually collected and stored in a distributed and privacy-sensitive manner. For instance, the multimedia data in smartphones, the sensor data collected in vehicles, and the examination data and diagnostic records of patients in different hospitals. These raise the problems such as data amount limitation and privacy concerns for traditional centralized multimodal machine learning [1].

To overcome these problems, federated learning [2] provides a paradigm to promote knowledge discovery of multimodal data under distributed scenarios. This paradigm allows the distributed clients collaboratively train a better-performed global model without sharing their local data. However, the majority of the previous work focuses on the unimodal setting where all the clients in the federated system hold the same data modality, as shown in Figure 1. Among those studies, statistical heterogeneity [3], *a.k.a.* the non-IID challenge, caused by the skew of label, features, and data quantity among clients, is one of the most critical challenges that attracts much attention [4–8]. However, multimodal federated learning further introduces modality heterogeneity, which leads to significant differences in model structures, local tasks, and parameter spaces among clients, thereby exposing substantial limitations of traditional uni-modal algorithms.

The federated systems trained with multimodal data are intuitively more powerful and insightful compared to unimodal ones [1]. We define the modality types held by the clients as their modality combinations, which will determine the local tasks they perform. If two clients hold the same or similar modality combinations (e.g., both image and text data), they will have a smaller semantic gap and task gap. In other words, the more congruent modality combinations the clients hold, the less heterogeneous the modality distribution of the system is.

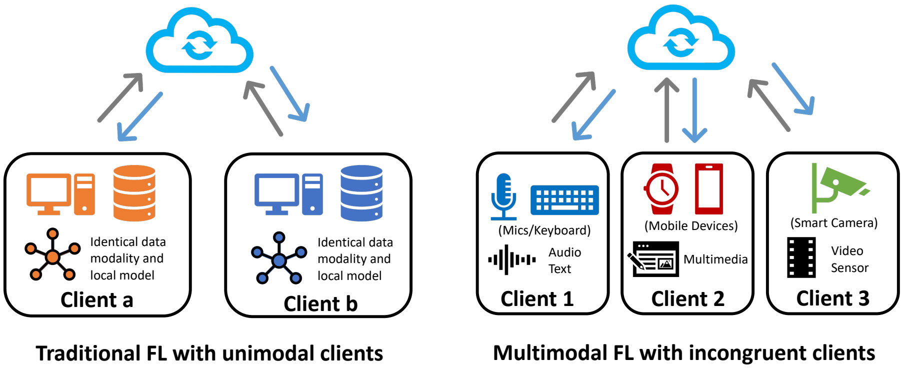
Identical data modality and local model
Client a
Identical data modality and local model
Client b
Traditional FL with unimodal clients
(Mics/Keyboard)
Audio
Text
Client 1
(Mobile Devices)
Multimedia
Client 2
(Smart Camera)
Video
Sensor
Client 3
Multimodal FL with incongruent clients

Figure 1. Illustration of traditional unimodal FL v.s. multimodal FL.

Based on the congruence of modality distribution, the MFL can be categorized into two categories: congruent MFL and incongruent MFL, as depicted in Figure 2. In the congruent MFL, the clients hold similar or the same local modality combinations, where the horizontal FL is the typical setting of this type. The majority of existing MFL work [9–12] also focuses on this federated setting, where all the clients hold the same input modality categories and feature space but differ on the sample space. In [10], the authors propose a multimodal federated learning framework for multimodal activity recognition with an early fusion approach via local co-attention. The authors in [12] provide a detailed analysis of the convergence problem of MFL with late fusion methods under the non-IID setting. In the healthcare domain [13–15], the congruent MFL has shown great application value by providing diagnosis assistance with distributed digital health data.

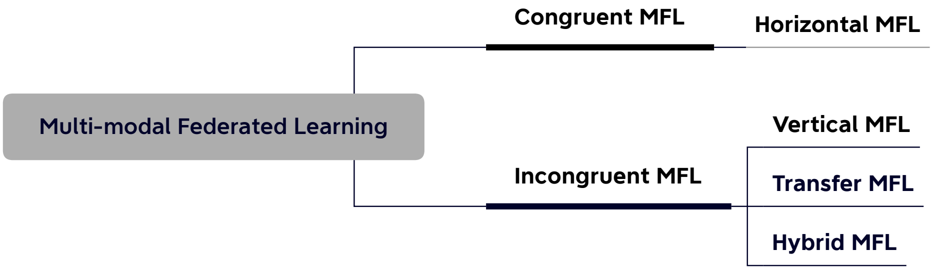

Figure 2. Taxonomy of multimodal federated learning (MFL).

For the incongruent MFL, the clients usually hold unique or partially overlapped data modality combinations, which makes the federated optimization and model aggregation more challenging. This category contains vertical multimodal federated learning (VMFL), multimodal federated transfer learning (MFTL), and hybrid multimodal federated learning (Hybrid-MFL). In VMFL, the clients hold different input modalities and feature spaces, but all the data samples are in the same space. In [16], the authors assume that each client only contains one specific modality and, correspondingly, propose FDARN, a five-module framework, for cross-modal federated human activity recognition (CMF-HAR). For MFTL, the clients mainly differ on feature spaces (e.g., photographic images and cartoon images) and sample ID spaces. For instance, in [17], the authors propose a fine-grained representation block named aimNet. They evaluate their methods on different FL settings, including the transfer setting between two different vision-language tasks.

Hybrid-MFL is a more challenging setting, where the data relationships among the clients can not be appropriately described by any of the above three settings alone. The clients in a hybrid setting can hold different local data varying on both modality categories and quantities. Given M modalities in a federated system, the theoretical client types are $2^M - 1$, including both unimodal and multimodal clients. [18] states a significant challenge for hybrid-MFL, which is modality incongruity, where the unique modality combination among the clients enlarges the heterogeneity. They proposed FedMSplit for multi-task learning in the hybrid-MFL setting, with a graph-based attention mechanism to extract the client relationship for aggregation.

Consequently, we provide a perspective on exploring the multimodal data in federated learning. In Section 2, we summarize the three popular aspects for mitigating the statistical heterogeneity in uni-modal federated learning systems. In Section 3, we give preliminaries and a formal definition of multimodal federated learning. In Section 4, we categorize multimodal federated learning into four types based on the input modalities of the clients. We introduce the common tasks and benchmarks for MFL in Section 5 and Section 6 separately. Section 7 identifies the challenges and promising directions, as well as the potential scenarios.

2. Federated Learning for Unimodal Heterogeneity

To mitigate the optimization divergence caused by heterogeneity challenges and increase the system robustness under the non-IID setting, existing unimodal methods are mainly proposed from three aspects: federated convergence optimization, personalized federated learning, and federated multi-task learning.

2.1. Federated Convergence Optimization

From the convergence optimization aspect, Zhao et al. in [4] investigate the under-performance problem of FedAvg [2] under the non-IID setting. They define and indicate that weight divergence causes the performance reduction issue and give a further analysis of that. In [19], the authors discover that the non-IID data will cause the problem of client drift in the uploading process and affect the system convergence rates. FedProx [20] modifies the local objective of each client, adding a proximal term to it. In [21], the authors propose SCAFFOLD to mitigate the client drifts between the local models and the global model. In [22], FedBVA decomposes the aggregation error into bias and variance for collaborative adversarial training in order to improve convergence and performance. In [23], the authors introduce reinforcement learning into the federated global update stage, where the server will dynamically select a subset of clients with the highest rewards to mitigate the heterogeneity challenge. The mutual target of those studies is to mitigate the divergence or drift problem during the global updating so that the framework can achieve a more generalized global model.

2.2. Personalized Federated Learning

Personalized federated learning [24] is a research topic proposed to handle the statistical heterogeneity challenge (i.e., the non-IID setting) from another aspect, where many existing methods aim to provide each client with personalized adaptation instead of a one-fit-all global solution. Ruan and Joe-Wong in [25] propose FedSoft, which reduces the clients' workload using proximal updates and brings both personalized local models and global cluster models. PerFedAvg [26] adapts meta-learning into the FL framework, where it treats the global model as a meta-model to give a few-shot adaptation for each client. In [27], the authors add Moreau envelopes as a regularization term in the local loss functions to help achieve personalized model optimization. In [28], the authors reassemble models and select the most fitted personalized models for clients by calculating the similarity.

2.3. Federated Multi-task Learning

Instead of seeking a general solution, federated multi-task learning [29] aims to find personalized models for each client by utilizing the similarity among them. MOCHA [29] is the first proposed federated multi-task learning framework for convex models. Corinzia et al. in [30] propose VIRTUAL, which uses approximated variational inference and simulates the federated system with a star-shaped Bayesian network. Marfoq et al. in [31] use federated expectation-maximization algorithm to solve the multi-task problem with a mixture of unknown underlying distributions.

The existing methods have shown great achievement in alleviating the statistical heterogeneity challenge via global optimization or learning personalized local models. However, as the heterogeneity among different client distributions diverges, the unified global model may not exist. Especially in the multimodal setting, federated clients could have different local model structures due to the modality heterogeneity, which makes the general solution more unfeasible. The local model personalization could be inapplicable either, due to the difference in both feature space and parameter space for MFL. In such a case, exploring multimodal data under the federated learning paradigm is not a direct combination of the existing techniques. Instead, the introduction of modality heterogeneity among the clients brings unique challenges and makes the existing challenge even more demanding.

3. Preliminaries of Multimodal Federated Learning

Compared to unimodal federated learning, we define multimodal federated learning as the federated systems containing at least two data modalities among all the local data sets. In the following, we will formally define multimodal federated learning using the multimodal classification task as an example.

In a multimodal federated learning system, given K clients and a modality category number M, where $K \geq 2$ and $M \geq 2$. Each client k is assumed to have access to a local data set $\mathbf{D}_k$, which contains a group of data sample IDs. The size of the data set is determined by the number of sample IDs, i.e., $|\mathbf{D}_k|$; M_k represents the total data modality number in client k and $M_k \in [1, M]$. If $M_k = 1$, client k is called *unimodal client*, and it is a *multimodal client* if $M_k > 1$.

To identify the types of clients based on their local data modalities, we define the local multimodal dataset $\mathbf{D}_k$ of an arbitrary client k as:

$$\mathbf{D}_k = \{(x_k^{m_1}, x_k^{m_2}, \dots, x_k^{m_{M_k}}, y_k)_i\}_{i=1}^{|\mathbf{D}_k|} \tag{1}$$

where x_k^m represents a data sample of m-modality in client k. The i-th data sample of the k-th local dataset is $\mathbf{X}_k(i) = (x_k^{m_1}, x_k^{m_2}, \dots, x_k^{m_{M_k}})_i$. The modality combination of this local set is defined as $\mathcal{X}_k = (m_1, m_2, \dots, m_{M_k})$. As an example, for client a containing both image and text data, its modality combination is $\mathcal{X}_a = (image, text)$ and its i-th local data sample is $\mathbf{X}_a(i) = (x_a^{image}, x_a^{text})_i$. Therefore, its modality number M_a is 2.

In a communication round t, the local model θ_k^t of client k can be updated by local training process via stochastic gradient descent(SGD):

$$\theta_k^{t+1} = \theta_k^t - \mu \nabla \mathcal{L}_k(\mathbf{X}_k, \theta_k^t, \boldsymbol{y}_k) \tag{2}$$

where μ is the learning rate of the local training process;$\mathbf{X}_k$ is the corresponding local multimodal data; $\mathcal{L}_k$ represents the total loss function of client k with multimodal input data $\mathbf{X}_k$; θ_k^t is the local model of client k parameters at communication round t.

Multiple modalities can make different contributions to the final loss affected by the problem context, data quality, and downstream tasks. For instance, in an image-text pair classification task, we may set a higher weight for the loss computed from image data and a lower one for text data. Therefore, given the input $\mathbf{X}_k(i)$, the total loss $\mathcal{L}_k$ is defined as:

$$\mathcal{L}_k(\mathbf{X}_k(i), \theta_k^t, \boldsymbol{y}_k(i)) = \sum_{j=1}^{M_k} \varphi_k^{m_j} l_k^{m_j}(C_k(x_k^{m_j}; \theta_k^t), \boldsymbol{y}_k(i)) \tag{3}$$

Here $\varphi_k^{m_j}$ represents the sum weight of modality m_j; C_k is the local model of client k; $l_k^{m_j}$ is the loss function for modality m_j; $x_k^{m_j}$ is the input data of modality m_j.

Accordingly, we define the local training target as follows:

$$f_k = \frac{1}{|\mathbf{D}_k|} \sum_{i=1}^{|\mathbf{D}_k|} \mathcal{L}_k(\mathbf{X}_k(i), \theta_k, \boldsymbol{y}_k(i)) \tag{4}$$

Thus, the global optimization target is defined as:

$$\min_{\theta_G} F(\theta_G) = \sum_{k=1}^{K} \omega_k f_k(\theta_k) \tag{5}$$

where θ_G is the global model parameters; ω_k is the global aggregation weight for client k; K is the total number of the clients.

4. Taxonomy of Multimodal Federated Learning

In [32], federated learning is categorized into horizontal federated learning, vertical federated learning, and federated transfer learning based on the data distribution characteristics. This previously proposed three-way division can clearly identify different categories of federated learning settings in unimodal scenarios.

However, it is not appropriate to directly adapt this categorization into multimodal federated learning. As the modality number of the local data expands, the distribution characteristics of the clients become more divergent. It is not illustrative enough to describe the data distribution relationship in such a way, especially when the clients contain different combinations of data modalities, i.e., modality incongruity challenge proposed by [18].

For instance, in a mental health prediction task, three mobile users participating in the federated systems may hold different preferences for digital APPs. As a result, their local datasets differ from data modalities and samples. There could exist the same data modalities such as screen time, typing history, and common sensor data. The users can also hold their unique modalities, such as image, video, audio, and APP data. In such a case, it is inappropriate to describe this federated system with sole horizontal federated learning or federated transfer learning.

In the above-mentioned case, the relationships among clients could be decomposed into modality-wise levels. To describe the combined relationships as such, we extend the taxonomy by introducing the *hybrid multimodal federated learning*. Thus, multimodal federated learning could be divided into four settings, which can be summarized into two categories based on the modality congruence, i.e., congruent MFL and incongruent MFL, as shown in Figure 2. Congruent MFL mainly covers horizontal settings, where all the clients hold the same modality set. Incongruent MFL, including vertical, transfer, and hybrid, allows clients to have totally different or partially overlapped modality sets. We introduce the four categories in the below subsections and summarize them in Table 1.

Table 1. Taxonomy of multimodal federated learning based on input modalities.

Taxonomy	Characteristics	Example	Related Work
Horizontal	Same input modalities, same feature space, different sample ID space.	Mobile phone users who use similar apps.	[9,14,15,17,33–36]
Vertical	Different input modalities, different feature space, same sample ID space.	IoT devices from different companies owned by a single user.	[13,16,37]
Transfer	Different input modalities, different feature space, different sample ID space.	The federated collaboration between the healthcare centers with different socioeconomic conditions and locations	[17,38,39]
Hybrid	Mixed combinations of different, partially different, or even same in input modalities, feature space, and sample ID space.	Mental health prediction task with diverse mobile users	[12,15,18,40]

4.1. Horizontal Multimodal Federated Learning

Similar to the unimodal horizontal setting, horizontal multimodal federated learning is defined as multimodal distributed systems where all the clients share the same modality combinations and the same data feature space but differ in sample IDs.

Definition 1 (Horizontal Multimodal Federated Learning)**.** *Given a client set $\mathcal{N}$ and modality set $\mathcal{M}$ in a federated system, the system is called Horizontal Multimodal Federated Learning, if for $\forall a, b \in \mathcal{N}$, they hold* *same modality set, i.e. $|M_a| = |M_b|$ and $\mathcal{X}_a = \mathcal{X}_b$. Here $|M_k|$ denotes the total numbers of modality types for client k and $\mathcal{X}_k$ is the modality combination set.*

For instance, in Figure 3 (left), two mobile users, $\mathbf{X}_a$ and $\mathbf{X}_b$, with the same APP usage patterns can hold both image and text data (denoted as x^{image} and x^{text} modalities) on their devices, as shown in Figure 3 (left). With the same data modalities locally, the two clients have inputs that are the same in modality combination but different in sample IDs, mathematically defined as follows:

$$\mathbf{X}_a = \{(x_a^{image}, x_a^{text}, y_a)_i\}_{i=1}^{|\mathbf{D}_a|}, \mathbf{X}_b = \{(x_b^{image}, x_b^{text}, y_b)_j\}_{j=1}^{|\mathbf{D}_b|}, \tag{6}$$

where $(x_a^{image}, x_a^{text}, y)_i$ denotes the i-th data sample of user a with two modalities *image* and *text* and the corresponding data label y. $|\mathbf{D}_a|$ represents the number of data samples.

In order to tackle the horizontal multimodal federated challenge in IoT systems, Zhao et al. in [33] propose a generation-based method, which utilizes autoencoder as the feature extractor to support the downstream classifier in the server side. The authors also validate the effectiveness of their method in the modality missing challenge, where some clients only have part of shared data modalities in the horizontal federation. In [9], the authors use an ensemble of local and global models to reduce both data variance and device variance in the federated system.

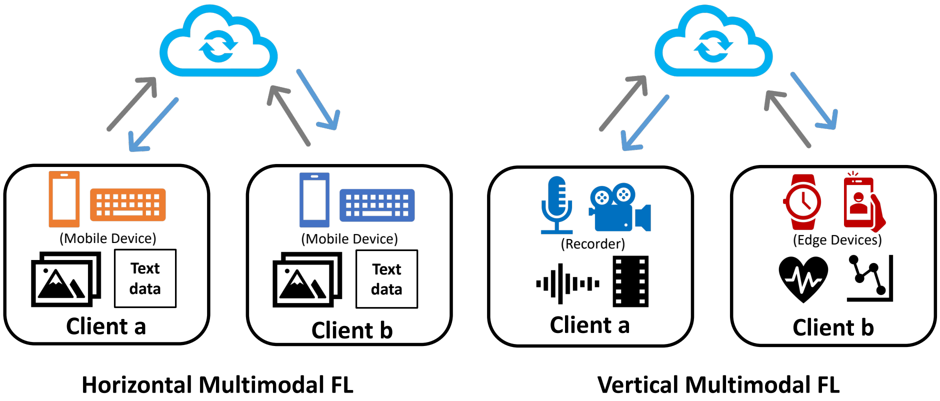

Figure 3. The illustration of Horizontal Multimodal Federated Learning and Vertical Multimodal Federated Learning. *Left:* Horizontal Multimodal Federated Learning contains two clients. Both hold image and text data. *Right:* Vertical Multimodal Federated Learning example contains two clients with exclusive modalities. Client a has audio and video data, while client b holds heat rate and acceleration sensor data.

4.2. Vertical Multimodal Federated Learning

Vertical multimodal federated learning defines a system with multiple unique data modality combinations held by different clients, where clients differ in the feature space but share the same sample ID set. These modality combinations are also exclusive without overlap on data modalities. Those modalities could be aligned well in either spatial or temporal relationships.

Definition 2 (Vertical Multimodal Federated Learning)**.** *Given a client set $\mathcal{N}$ and modality set $\mathcal{M}$ in a federated system, the system is defined as Vertical Multimodal Federated Learning, if for $\forall a, b \in \mathcal{N}$, they hold totally different modality combinations, while connected by sample IDs, i.e., $\mathcal{X}_a \cap \mathcal{X}_b = \emptyset$ and $\mathbf{D}_a = \mathbf{D}_b$.*

For instance, in the human activity recognition task, a user may own multiple devices which collect different data modalities due to the divergence of sensor category, as shown in Figure 3 (right). In a two-device case, the local data sets of the devices could be defined as:

$$\mathbf{X}_a = \{(x_a^{video}, x_a^{audio}, y_a)_i\}_{i=1}^{|D|}, \mathbf{X}_b = \{(x_b^{heart\_rate}, x_b^{acceleration}, y_b)_i\}_{i=1}^{|D|}, \tag{7}$$

where client a holds modality *video* and modality *audio*, and modality *heartratesensor* and *accelerationsensor* for client b. Unlike the horizontal scenario, the two clients could share the same sample ID set $\mathbf{D}$.

In [16], the authors propose the feature-disentangled activity recognition network (FDARN) for the cross-modal federated human activity recognition task. With five adversarial training modules, the proposed method captures both the modality-agnostic features and modality-specific discriminative characteristics of each client to achieve a better performance than existing personalized federated learning methods. Notably, each client holds a single modality data set that may differ from group to group in their experiments.

4.3. Multimodal Federated Transfer Learning

Multimodal federated transfer learning supposes that the clients in the federation have different input modalities and sample ID sets. Different clients are allowed to have some overlap of the modality combinations while differing from the feature space. In other words, the clients may conduct different local tasks, such as VQA and image captioning for vision-language clients.

Definition 3 (Multimodal Federated Transfer Learning)**.** *Given a client set $\mathcal{N}$ and modality set $\mathcal{M}$ in a federated system, the system is defined as Multimodal Federated Transfer Learning, if for $\forall a, b \in \mathcal{N}$, they hold different modality combinations and sample ID, $\mathcal{X}_a \cap \mathcal{X}_b = \emptyset$ and $D_a \neq D_b$.*

As shown in Figure 4 (left), considering a federated learning example between different hospitals, a hospital located in a developed area usually takes advance in equipment compared to one in a rural area. In addition, due to the locations, the two hospitals may receive different patient groups. These may result in differences in their data modalities and sample ID sets stored in their databases. For a two-party multimodal federated transfer learning system, the input modalities of the clients can be represented as:

$$\mathbf{X}_a = \{(x_a^{MRI}, x_a^{PET}, y_a)_i\}_{i=1}^{|\mathbf{D}_a|}, \mathbf{X}_b = \{(x_b^{MRI}, x_b^{CT}, y_b)_j\}_{j=1}^{|\mathbf{D}_b|}, \tag{8}$$

where the two clients differ in both local data modalities and sample ID sets. However, since CT, MRI, and PET scans are all medical image techniques for diagnosis. The rich knowledge and model advantages could be shared between the clients, which forms a typical multimodal federated transfer learning setting.

Liu et al. propose aimNet to generate fine-grained image representations and improve the performance on various vision-and-language grounding problems under federated settings. They validate their methods in horizontal, vertical, and transfer multimodal federated learning settings to show their superiority.

4.4. Hybrid Multimodal Federated Learning

Hybrid multimodal federated learning is defined as a federated system where all the clients have incongruent data modalities in their local sets. The modality combination of each client is unique in the system, which varies from modality quantity and category. Both unimodal and multimodal clients can exist in the system.

Definition 4 (Hybrid Multimodal Federated Learning)**.** *Given a client set $\mathcal{N}$ and modality set $\mathcal{M}$ in a federated system, the system is defined as Hybrid Multimodal Federated Learning, if there exist at least two basic relationships from horizontal, vertical, transfer or both unimodal and multimodal clients. There are utmost $2^M - 1$ types of clients in a hybrid federated system.*

In Figure 4 (right), we may take the mental health prediction task at the beginning of the section as an example of hybrid MFL, where three mobile users share a horizontal-related screen time(ST) and hold a unique data modality, respectively. The input modalities of this example are:

$$\mathbf{X}_a = \{(x_a^{ST}, x_a^{image}, y_a)_i\}_{i=1}^{|D_a|}, \mathbf{X}_b = \{(x_b^{ST}, x_b^{video}, y_b)_j\}_{j=1}^{|D_b|}, \mathbf{X}_c = \{(x_c^{ST}, x_c^{audio}, y_c)_k\}_{k=1}^{|D_c|}. \tag{9}$$

The client category could be various in a hybrid setting. In a bi-modal federated system, there could be totally three kinds of clients; for a tri-modal federated system, the client category could rise to seven by different modality numbers and combinations. The relationships among the clients in a hybrid MFL system could be described at the modality level.

Chen and Zhang in [18] propose FedMSplit, a dynamic and multi-view graph structure aiming to solve the modality incongruity challenges in hybrid MFL setting. The novel modality incongruity problem in MFL is a significant challenge in the scope of hybrid MFL. In [40], the authors propose a general multimodal model that works on both multi-task and transfer learning for high-modality (a large set of diverse modalities) and partially-observable (each task only defined on a small subset of modalities) scenarios. This indicates the recent research trend to design more general-purpose multimodal models and reveals the importance of exploring hybrid MFL, the most challenging and complex multimodal federated scenario.

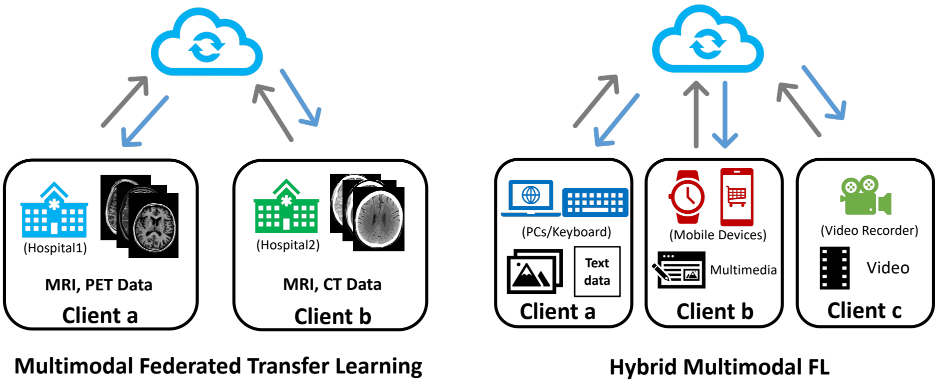

Figure 4. The illustration of Multimodal Federated Transfer Learning and Hybrid Multimodal Federated Learning. *Left:* Multimodal Federated Transfer Learning contains two hospitals as clients. One holds MRI and PET data, the other holds MRI and CT data. *Right:* Hybrid Multimodal Federated Learning example contains three clients with different modality combinations. The system contains both unimodal and multimodal clients.

5. Tasks for Multimodal Federated Learning

Multimodal federated learning (MFL) offers many advantages, such as privacy preservation and addressing the data silo problem. However, it also faces limitations such as communication costs, data heterogeneity, and hardware disparities compared to centralized multimodal learning. Therefore, in addition to the unique challenges of modal heterogeneity, the original multimodal learning tasks become more challenging when performed within a federated learning framework. In this section, we will discuss several representative MFL application tasks.

5.1. Vision Language Interaction

Visual and language data are widely present in data centers and edge devices, making visual-language interaction an important task in MFL. Specifically, a federated learning system targeting visual and language data should be capable of handling diverse and complex visual-language learning tasks, including visual question answering, visual reasoning, image captioning, image-text retrieval, and text-to-image generation. In the context of local training on client devices, the system needs to achieve efficient and robust multimodal matching and cross-modal interactions. On the one

doi:10.20944/preprints202307.1420.v1

hand, due to constraints imposed by client hardware and communication requirements, lightweight and high-performance characteristics are desired in MFL approaches. Integrating state-of-the-art pre-trained large-scale models from the forefront of federated learning and visual-language learning becomes a promising research direction. On the other hand, heterogeneity in data, particularly in terms of label and modality, often leads to differences in model architectures and tasks among clients. These task gaps and semantic gaps can negatively impact global aggregation on the server side, posing challenges for achieving convergent global optimization.

Several pioneering studies have explored the field of MFL in the context of visual-language tasks. In [17], the authors propose aimNet and evaluate it under horizontal FL, vertical FL, and Federated Transfer Learning (FTL) settings when the clients are conducting either the VQA task or image captioning task. CreamFL [38] utilizes contrastive learning for ensembling the uploaded heterogeneous local models based on their output representations. The CreamFL [38] allows both uni-modal and multimodal vision-language tasks in federated systems. pFedPrompt [35] adapts prompt training manner to leverage the large foundation model into federated learning systems to connect vision and language data. FedCMR [11] explores the federated cross-modal retrieval task and mitigates the representation space gap via weighted aggregation based on local data amount and category number.

5.2. Human Activity Recognition

The wireless distributed sensor system, such as the IoT system, is a signification application scenario for MFL, where multiple sensors provide consistent observations of the same object or event. Human activity recognition (HAR) is one of the most representative tasks in this setting, due to the privacy preservation requirement.

The data partition method for the HAR task in existing MFL works has two types, client-as-device and client-as-sensor. The former one is represented by MMFed [10], which equally divided the multimodal data for each client. The local co-attention mechanism then performs multimodal fusion. Zhao et al. conduct their experiment by giving each client only a single type of modality [33]. The local network is divided into five modules for either modality-wise aggregation for clients with the same modality or general aggregation for all clients. However, the modality distribution or data partition way could be more diverse according to the hardware deployment and environmental factors.

5.3. Emotion Recognition

Emotion recognition plays a crucial role in improving social well-being and enhancing societal vitality. The multimodal data generated during the use of mobile phones often provide valuable insights into identifying users who may have underlying mental health issues. Effective emotion recognition algorithms can target specific users to enhance their experience and prevent the occurrence of negative events such as suicide and depression. However, multimedia data associated with user emotions is highly privacy-sensitive, including chat records and personal photos. In this context, the MFL framework offers the capability of efficient collaborative training while ensuring privacy protection. Therefore, emotion recognition undoubtedly holds a significant position within the realm of MFL.

There are several MFL works that investigate the emotion recognition task in the vertical and hybrid MFL setting. In [37], each client in the system contains only one modality and the uni-modal encoders are trained on the local side. The proposed hierarchical aggregation method aggregates the encoders based on the modality type held by the clients and utilizes an attention-based method to align the decoder weights regardless of the data modality. The FedMSplit approach [18] utilizes a dynamic and multi-view graph structure to flexibly capture the correlations among client models in a multimodal setting. Liang et al. in [41] propose a decentralized privacy-preserving representation learning method that uses multimodal behavior markers to predict users' daily moods and identify the early risks of suicide.

5.4. Healthcare

Numerous healthcare centers and hospitals have accumulated vast amounts of multimodal data during patient consultations and treatments, including X-ray images, CT scans, physician diagnoses, and physiological measurements of patients. These multimodal data are typically tightly linked to patient identifiers and require stringent privacy protection measures. As a result, these healthcare institutions have formed isolated data islands, impeding direct collaboration in terms of cooperative training and data sharing through open databases. This presents a series of crucial challenges within the realm of multimodal federated learning, encompassing tasks such as AI-assisted diagnosis, medical image analysis, and laboratory report generation.

Some works in the field of healthcare have explored multimodal federated learning, often assuming that all institutions have the same set of modalities, referred to as horizontal MFL, or that each institution possesses only a single modality, known as vertical MFL. Agbley et al. in [14] apply federated learning on the prediction of melanoma and obtained performance that is on-par with the centralized training result. FedNorm [15] performs modality-based normalization techniques to enhance liver segmentation and trains over uni-modal clients holding CT and MRI data separately. Qayyum et al. utilize the cluster federated learning for automatic diagnosis of COVID-19 [13]. Each cluster contains the healthcare entities that contain the same modality, such as X-ray and Ultrasound.

6. Benchmarks for Multimodal Federated Learning

As discussed in Section 5, multimodal federated learning exhibits numerous broad application scenarios and tasks. However, the benchmarking of MFL frameworks specifically designed for testing and executing MFL tasks is still in its exploratory stage. Therefore, in this section, we present a series of benchmark datasets suitable for multimodal federated learning to facilitate further research endeavors.

6.1. Vision Language Datasets

Caltech-UCSD Birds-200-2011 (CUB-200-2011). CUB-200-2011 [42] is one of the most widely used fine-grained categorization datasets. It contains 11,788 images of 200 subcategories belonging to birds. Each image has its own annotations for identification, which include one subcategory label, one bounding box, 15 part locations, and 312 binary attributes. Reed et al. expand the dataset by providing ten fine-grained text description sentences for each image [43]. The sentences are collected through the Amazon Mechanical Turk (AMT) platform and have a minimum length of 10 words, without exposing the label and action information.

Oxford 102 Flower (102 Category Flower Dataset). Oxford 102 Flower [44] is a fine-grained classification dataset, which collects 102 categories of flowers that commonly occur in the United Kingdom. Each category contains 40 to 258 images. There are 10 text descriptions for each image.

UPMC Food-101. The Food-101 [45] is a noisy multimodal classification dataset that contains both images and paired captions of 101 food categories. Each category has 750 training and 250 testing images. There are a total of 101,000 images, each paired with one caption. However, the labels and captions of the training set contain some noise and may leak the label information. The testing set has been manually cleaned.

Microsoft Common Objects in Context (MS COCO). The MS COCO dataset [46] is a comprehensive dataset used for various tasks such as object detection, segmentation, key-point detection, captioning, stuff image segmentation, panoptic segmentation, and dense pose estimation. It comprises a total of 328K images. The dataset provides detailed annotations for object detection (bounding boxes and segmentation masks), captioning, keypoint detection, stuff image segmentation, panoptic segmentation, and dense pose estimation. Note that the dense pose annotations are only available for training and validation images, totaling more than 39,000 images and 56,000 person instances.

Flickr30k. The Flickr30k dataset [47] comprises 31,000 images sourced from Flickr, accompanied by 5 reference sentences per image generated by human annotators. Additionally, we have constructed an image caption corpus consisting of 158,915 crowd-sourced captions describing 31,783 images. This updated collection of images and captions primarily focuses on individuals participating in routine activities and events.

6.2. Human Activity Recognition Datasets

NTURGB+D120. The NTU RGB+D 120 dataset [48] is a comprehensive collection specifically designed for RGB+D human action recognition. It comprises a vast amount of data sourced from 106 unique subjects, encompassing over 114 thousand video samples and 8 million frames. This dataset encompasses a wide range of 120 distinct action classes, encompassing various activities that are part of daily routines, mutual interactions, and health-related actions. It serves as a valuable resource for research and development in the field of human action recognition, facilitating advancements in computer vision, machine learning, and artificial intelligence applications.

Epic-Kitchens-100. EPIC-KITCHENS-100 [49] is a large-scale dataset focusing on first-person (egocentric) vision. It features multi-faceted audio-visual recordings of individuals' daily kitchen activities captured in their homes using head-mounted cameras. The dataset, comprising 45 kitchens across four cities, offers diverse environmental contexts. With 100 hours of Full HD footage and 20 million frames, it provides a rich visual experience for analysis. The annotations, obtained through a unique 'Pause-and-Talk' narration interface, enhance content understanding. The dataset includes $90,000$ action segments, $20,000$ unique narrations, and supports multiple languages, facilitating cross-cultural studies. It covers a wide range of activities classified into 97 verb classes and 300 noun classes, enabling fine-grained analysis within the kitchen context.

Stanford-ECM. Stanford-ECM [50] is an egocentric multimodal dataset containing approximately 27 hours of egocentric video recordings accompanied by heart rate and acceleration data. The video lengths vary from 3 minutes to around 51 minutes, ensuring a diverse range of content. The videos were captured using a mobile phone at $720x1280$ resolution and 30 fps, while the triaxial acceleration was recorded at 30Hz. A wrist-worn heart rate sensor captured heart rate readings every 5 seconds, and the phone and heart rate monitor were synchronized via Bluetooth. All data was stored in the phone's storage, with any gaps in heart rate data filled using piecewise cubic polynomial interpolation. The data was meticulously aligned to the millisecond level at a frequency of 30 Hz, ensuring precise synchronization across the modalities.

mHealth (Mobile Health). The mHealth (Mobile Health) dataset [51] is a collection of body motion and vital signs recordings from ten volunteers engaging in various physical activities. The dataset includes sensors placed on the chest, right wrist, and left ankle to measure acceleration, rate of turn, and magnetic field orientation across different body parts. Additionally, the chest sensor provides 2-lead ECG measurements, allowing for potential applications in basic heart monitoring, arrhythmia detection, and analysis of exercise effects on the ECG. Overall, the dataset consists of 12 activities performed by 10 subjects, with three sensor devices utilized for data collection.

6.3. Emotion Recognition Datasets

Interactive Emotional Dyadic Motion Capture (IEMOCAP). The IEMOCAP database [52] is a multimodal and multispeaker dataset designed for studying emotional expressions. It encompasses around 12 hours of audiovisual data, including video recordings, speech, facial motion capture, and text transcriptions. The database features dyadic sessions in which actors engage in improvised or scripted scenarios carefully crafted to elicit emotional responses. Multiple annotators have labeled the IEMOCAP database with categorical labels like anger, happiness, sadness, neutrality, as well as dimensional labels such as valence, activation, and dominance.

Multimodal Corpus of Sentiment Intensity (CMU-MOSI). The CMU-MOSI dataset [53,54] consists of 2199 opinion video clips, each annotated with sentiment values ranging from -3 to 3. The dataset includes detailed annotations for subjectivity, sentiment intensity, visual features annotated per frame and per opinion, as well as audio features annotated per millisecond.

CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI). The CMU-MOSEI dataset [53,54] is a multimodal data collection for analyzing sentiment and emotion in opinionated text and speech. It contains a total of 23, 453 video segments extracted from 1, 000 YouTube videos, with each segment accompanied by transcriptions, audio, and visual features. The dataset provides rich annotations for sentiment, emotion, and intensity, allowing researchers to explore the interplay between language, speech, and visual cues in expressing opinions and emotions.

6.4. Healthcare Datasets

MIMIC-IV. The Medical Information Mart for Intensive Care (MIMIC-IV) dataset [55] is a comprehensive and widely used database that provides detailed clinical data from patients admitted to intensive care units (ICUs). MIMIC-IV offers an expanded collection of de-identified electronic health records (EHRs) from diverse healthcare institutions. It contains a wealth of information, including vital signs, laboratory results, medications, procedures, diagnoses, and patient demographics. The dataset is invaluable for conducting research and developing algorithms and models related to critical care medicine, clinical decision-making, and healthcare analytics.

MIMIC-CXR. The MIMIC Chest X-ray (MIMIC-CXR) database [56] is a large, publicly available collection of de-identified chest radiographs in DICOM format accompanied by corresponding free-text radiology reports. It comprises 377, 110 images from 227, 835 radiographic studies conducted at the Beth Israel Deaconess Medical Center in Boston, MA. The dataset adheres to the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Safe Harbor requirements, ensuring the removal of protected health information (PHI). Its purpose is to facilitate diverse medical research areas, including image analysis, natural language processing, and decision support, providing a valuable resource for advancing knowledge and innovation in the field of medicine.

Alzheimer's Disease Neuroimaging Initiative (ADNI). The Alzheimer's Disease Neuroimaging Initiative (ADNI) database [57] is a comprehensive and widely used collection of data aimed at advancing research in Alzheimer's disease and related neurodegenerative disorders. ADNI consists of clinical, genetic, imaging, and biomarker data gathered from participants across multiple sites in the United States and Canada. The dataset includes various modalities such as magnetic resonance imaging (MRI), positron emission tomography (PET), cerebrospinal fluid (CSF) biomarkers, and cognitive assessments.

6.5. Multi-sensor Datasets

ModelNet40. The ModelNet40 dataset [58] comprises 3D synthetic object point clouds, making it a highly utilized benchmark in point cloud analysis due to its diverse categories, precise shapes, and well-organized dataset. In its original form, ModelNet40 encompasses 12, 311 CAD-generated meshes representing 40 categories, including objects like airplanes, cars, plants, and lamps. Out of these, 9, 843 meshes are designated for training purposes, while the remaining 2, 468 meshes are set aside for testing. The corresponding point cloud data points are uniformly sampled from the surfaces of these meshes and subsequently preprocessed by repositioning them to the origin and scaling them to fit within a unit sphere.

Vehicle Sensor. The Vehicle Sensor dataset [59] is proposed for the vehicle type classification task in wireless distributed sensor networks (WDSN). The dataset consists of 23 road segmentation instances. Each instance has 50 acoustic and 50 seismic features.

6.6. Multi-task Dataset

FedMultimodal. FedMultimodal [60] is the first federated learning benchmark designed for multimodal learning, encompassing five key multimodal applications across ten well-known datasets, featuring a total of eight distinct modalities. This benchmark introduces a comprehensive FL pipeline, enabling a holistic modeling framework that covers data partitioning, feature extraction, FL benchmark algorithms, and model evaluation. In contrast to existing FL benchmarks, FedMultimodal offers a standardized methodology for evaluating FL's resilience in real-world multimodal scenarios, specifically addressing three prevalent data corruptions: missing modalities, missing labels, and erroneous labels.

7. Discussion

We introduce the potential directions and challenges of multimodal federated learning in this section. These challenges are non-exclusive, rather they root in one core factor, the data modality distribution in the federated learning system.

7.1. Modality Heterogeneity

The heterogeneity problem in unimodal FL is usually caused by the imbalance of data quantity and data label skews. While the introduction of modality distribution will further increase the complexity of the problem, the heterogeneity in both statistical distribution and modality distribution will affect the global convergence and performance of the federated system. In addition, most existing non-IID and personalized methods are questionable in their effectiveness in multimodal federated settings. Thus, innovative and effective solutions are expected to be proposed for this setting.

In the MFL setting, the modality heterogeneity challenge exists at both the client level and the system level. At the client level, the clients require efficient local representation learning to bridge the semantic gap [61,62] among the multimodal data. One of the most popular solutions for multimodal representation learning is to map the different modalities of data into a mutual latent space. Similar to centralized multimodal learning, how to adapt advanced knowledge from representation learning and centralized multimodal learning to design feature extractor modules for merging those local gaps in the local learning process is a vital topic.

At the system level, there exists a task gap among all the clients caused by the difference in modality combination, e.g., the modality types in local datasets. In centralized multimodal learning, representation learning usually transforms the different modalities into a common representation space via embedding operation. Comparatively, MFL divides the common space of the centralized scenario into N common subspaces, making the unifying of the embedding operation among all the clients difficult. A client maps the original multimodal data into embedding representations, which all exist in its unique common subspace. Due to the modality heterogeneity and the different local model tasks, those common subspaces differs from each other, resulting the task gaps that are difficult to bridge. For example, unimodal clients and multimodal clients could hold totally different parameter spaces and work on different feature spaces. Even if the clients hold the same modality combinations, the specific local tasks can be different, such as visual question answering and image captioning.

To solve the challenge, it raises the requirement for the new aggregation paradigm. The modality heterogeneity could result in more divergent gradients and even heterogeneous local model architectures. As Wang et al. prove that different modalities overfit and generalize at different rates, the one-fit-all global training strategy for MFL might not work since optimizing all the clients jointly can result in sub-optimal [63].

7.2. Modality Missing

Another significant challenge for MFL is modality missing, pointing to some clients suffering data quantity imbalance among different modalities of their local data sets. For instance, for a client that holds $1,000$ image-text pair data, 300 of the pairs might lose their image data and 200 of them do not have the text part. The absence of modalities poses challenges to both the structure and robustness of models. Many transformer-based models [64] will encounter significant performance degradation in such cases.

The modality missing happens frequently in realistic scenarios caused by hardware limitations, collection errors, and storage issues. To address the issue of modality missing, there are some strategies and techniques for centralized learning have been proposed. Some approaches [65,66] involve data imputation or reconstruction methods to fill in the missing modalities based on the available information. Others leverage multi-task learning or meta-learning [64] techniques to transfer knowledge from modalities with sufficient data to those with missing data. Additionally, there are efforts to design more robust and flexible models [67] that can effectively handle missing modalities without significant performance degradation. Observing these developments of addressing the modality missing in centralized learning, it is crucial to explore lightweight and data-efficient methods for MFL to tackle the missing challenge.

7.3. Data Complexity

Multimodal federated learning serves as a promising solution for the collaborative ML-based system among healthcare entities and medical-related institutions. On the one hand, medical research and patient diagnosis generate massive multimodal data, which is a great resource to boost the development of advanced ML methods. On the other hand, those healthcare data, such as electronic health records (EHR), X-ray, and CT scan images, are stored and managed in a privacy-sensitive manner. MFL enables the exploration of information within such complex data, facilitating collaborative training across healthcare and medical data silos and thereby delivering improved AI-assisted diagnosis, medical image analysis, report generation, and other related services.

Different from the application in IoT or multimedia domain, the healthcare data are more complex and diverse in both format and granularity. The heterogeneity of data is further amplified among healthcare institutions due to differences in medical equipment, diagnostic methods, and data management practices, making federated collaboration [68] more difficult. A great amount of medical-related MFL work is emerging. Cobbinah et al. in [69] provide an FL-based method for the prediction of the phase of Alzheimer's disease utilizing the MRI data from multi-center. In [13,14], the authors use MFL for AI-assisted disease diagnosis and achieve satisfying performance. FedNorm [15] explores the medical image segmentation in the FL setting.

7.4. Large-scale Pre-trained Model

Among the recent progress of multimodal learning, the inspiring performance of the large-scale pre-trained models [70–72] has lightened a promising future for solving broad machine learning tasks with a unified and effective solution. However, there still exist two main challenges known to block the universal deployment of those large-scale pre-trained models in federated learning systems.

On the one hand, the cost of building and training those large-scale pre-trained models could be extremely expensive and not affordable for most computing devices and data centers. On the other hand, massive training data is required to gather for effective training of large foundation models. In a federated learning system, the clients usually have limited hardware resources and communication bandwidth, which makes it impractical to train from scratch or even fine-tune large-scale foundation models in FL scenarios. On the other hand, utilizing traditional knowledge distillation methods to transfer knowledge from pre-trained large models within the framework of federated learning also faces limitations. This is due to the fact that the training data for pre-training large models are often massive and diverse, making it challenging for clients in a federated learning system to collect and store such data. These existing limitations will hinder the deployment and empowerment of large-scale foundation models on distributed learning frameworks represented by federated learning. The capability to integrate the large-scale pre-trained models in a lightweight way is expected for an effective federated multimodal learning framework.

To overcome these limitations, several works have been proposed to tackle the deployment of large-scale pre-trained models in multimodal federated learning. FedBERT Tian et al. [73] adapts split learning to achieve efficient distributed training for large-scale BERT [70] model. The head and mapping layers of the BERT are distributed to clients for local training and then aggregated on the server side. Tan et al. proposed Federated Prototype-wise Contrastive Learning (FedPCL) algorithm uses class prototypes and contrastive learning to share class-relevant knowledge among clients, demonstrating improved personalization and knowledge integration capabilities with a pre-trained backbone model. FedCLIP [39] adds an adapter module after the CLIP backbone to achieve efficient deployment of CLIP model [71] at federated clients. Some studies [35,75] utilize the idea of prompt training that aggregate the user consensus via learnable prompt and improve the users' characteristics in the visual domain. Improving the capability of integrating large-scale pre-trained models will greatly enhance the performance of the MFL systems.

7.5. Weak-supervised Learning

In realistic application scenarios, the collected multimodal data in data silos usually contain limited supervision signals such as labels or matching relationships among the modalities. The exploration of weak-supervised learning in MFL is another crucial topic, which also includes self-supervised learning and semi-supervised learning.

There are some uni-modal FL works that have investigated the field of self-supervised learning [76,77] and semi-supervised learning [7,78–80] techniques in FL. However, the heterogeneity of multimodal data will further challenge the robustness of the system with limited and noisy supervision signals. These directions hold great potential for advancing the field of multimodal federated learning by enabling models to learn from diverse and abundant but weakly labeled data sources, thus paving the way for improved performance and generalization in real-world applications.

8. Conclusion

In this paper, we identify a promising research topic, multimodal federated learning. We give an introduction to existing MFL methods and the motivation for utilizing distributed multi-modal data. Furthermore, we extend the federated learning categorization to multimodal scenarios and show the related tasks under different multimodal settings. Finally, we discuss the potential challenges and perspectives on future directions.

Author Contributions: Conceptualization, L.C. and F.M.; methodology, L.C.; software, L.C.; validation, L.C., J.W., Y.Z. and F.M.; formal analysis, L.C.; investigation, L.C.; resources, L.C.; writing—original draft preparation, L.C.; writing—review and editing, J.W., Y.Z. and F.M.; visualization, L.C.; supervision, F.M.; project administration, F.M.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We sincerely thank all anonymous reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence* **2018**, *41*, 423–443.
2. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. Artificial intelligence and statistics. PMLR, 2017, pp. 1273–1282.
3. Zhu, H.; Xu, J.; Liu, S.; Jin, Y. Federated learning on non-IID data: A survey. *465*, 371–390. doi:https://doi.org/10.1016/j.neucom.2021.07.098.
4. Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; Chandra, V. Federated Learning with Non-IID Data **2018**. doi:10.48550/ARXIV.1806.00582.
5. Sattler, F.; Wiedemann, S.; Müller, K.R.; Samek, W. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems* **2019**, *31*, 3400–3413.
6. Wang, H.; Kaplan, Z.; Niu, D.; Li, B. Optimizing federated learning on non-iid data with reinforcement learning. IEEE INFOCOM 2020-IEEE Conference on Computer Communications. IEEE, 2020, pp. 1698–1707.

7. 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. Proceedings of the 2023 SIAM International Conference on Data Mining (SDM). SIAM, 2023, pp. 496–504.
8. Wang, J.; Qian, C.; Cui, S.; Glass, L.; Ma, F. Towards federated covid-19 vaccine side effect prediction. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2022, pp. 437–452.
9. Liang, P.P.; Liu, T.; Ziyin, L.; Allen, N.B.; Auerbach, R.P.; Brent, D.; Salakhutdinov, R.; Morency, L.P. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523* **2020**.
10. Xiong, B.; Yang, X.; Qi, F.; Xu, C. A unified framework for multi-modal federated learning. *Neurocomputing* **2022**, *480*, 110–118.
11. Zong, L.; Xie, Q.; Zhou, J.; Wu, P.; Zhang, X.; Xu, B. FedCMR: Federated Cross-Modal Retrieval. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1672–1676.
12. Chen, S.; Li, B. Towards Optimal Multi-Modal Federated Learning on Non-IID Data with Hierarchical Gradient Blending. IEEE INFOCOM 2022-IEEE Conference on Computer Communications. IEEE, 2022, pp. 1469–1478.
13. Qayyum, A.; Ahmad, K.; Ahsan, M.A.; Al-Fuqaha, A.; Qadir, J. Collaborative federated learning for healthcare: Multi-modal covid-19 diagnosis at the edge. *arXiv preprint arXiv:2101.07511* **2021**.
14. Agbley, B.L.Y.; Li, J.; Haq, A.U.; Bankas, E.K.; Ahmad, S.; Agyemang, I.O.; Kulevome, D.; Ndiaye, W.D.; Cobbinah, B.; Latipova, S. Multimodal melanoma detection with federated learning. 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). IEEE, 2021, pp. 238–244.
15. Bernecker, T.; Peters, A.; Schlett, C.L.; Bamberg, F.; Theis, F.; Rueckert, D.; Weiß, J.; Albarqouni, S. FedNorm: Modality-Based Normalization in Federated Learning for Multi-Modal Liver Segmentation. *arXiv preprint arXiv:2205.11096* **2022**.
16. Yang, X.; Xiong, B.; Huang, Y.; Xu, C. Cross-Modal Federated Human Activity Recognition via Modality-Agnostic and Modality-Specific Representation Learning **2022**.
17. Liu, F.; Wu, X.; Ge, S.; Fan, W.; Zou, Y. Federated learning for vision-and-language grounding problems. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 34, pp. 11572–11579.
18. Chen, J.; Zhang, A. FedMSplit: Correlation-Adaptive Federated Multi-Task Learning across Multimodal Split Networks. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery: New York, NY, USA, 2022; KDD '22, p. 87–96. doi:10.1145/3534678.3539384.
19. Li, X.; Huang, K.; Yang, W.; Wang, S.; Zhang, Z. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189* **2019**.
20. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated Optimization in Heterogeneous Networks, 2018. doi:10.48550/ARXIV.1812.06127.
21. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; Suresh, A.T. Scaffold: Stochastic controlled averaging for federated learning. International Conference on Machine Learning. PMLR, 2020, pp. 5132–5143.
22. Zhou, Y.; Wu, J.; Wang, H.; He, J. Adversarial Robustness through Bias Variance Decomposition: A New Perspective for Federated Learning. Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022. ACM, 2022, pp. 2753–2762.
23. Wang, H.; Kaplan, Z.; Niu, D.; Li, B. Optimizing Federated Learning on Non-IID Data with Reinforcement Learning. IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, 2020, pp. 1698–1707. doi:10.1109/INFOCOM41043.2020.9155494.
24. Tan, A.Z.; Yu, H.; Cui, L.; Yang, Q. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems* **2022**.
25. Ruan, Y.; Joe-Wong, C. Fedsoft: Soft clustered federated learning with proximal local updating. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 8124–8131.
26. Fallah, A.; Mokhtari, A.; Ozdaglar, A. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948* **2020**.

27. T Dinh, C.; Tran, N.; Nguyen, J. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems* **2020**, *33*, 21394–21405.
28. Wang, J.; Cui, S.; Ma, F. FedLEGO: Enabling Heterogenous Model Cooperation via Brick Reassembly in Federated Learning. International Workshop on Federated Learning for Distributed Data Mining, 2023.
29. Smith, V.; Chiang, C.K.; Sanjabi, M.; Talwalkar, A.S. Federated multi-task learning. *Advances in neural information processing systems* **2017**, *30*.
30. Corinzia, L.; Beuret, A.; Buhmann, J.M. Variational federated multi-task learning. *arXiv preprint arXiv:1906.06268* **2019**.
31. Marfoq, O.; Neglia, G.; Bellet, A.; Kameni, L.; Vidal, R. Federated Multi-Task Learning under a Mixture of Distributions. Advances in Neural Information Processing Systems; Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; Vaughan, J.W., Eds. Curran Associates, Inc., 2021, Vol. 34, pp. 15434–15447.
32. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2019**, *10*, 1–19.
33. Zhao, Y.; Barnaghi, P.; Haddadi, H. Multimodal Federated Learning on IoT Data. 2022 IEEE/ACM Seventh International Conference on Internet-of-Things Design and Implementation (IoTDI). IEEE, 2022, pp. 43–54.
34. Zong, L.; Xie, Q.; Zhou, J.; Wu, P.; Zhang, X.; Xu, B. FedCMR: Federated Cross-Modal Retrieval. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; Association for Computing Machinery: New York, NY, USA, 2021; SIGIR '21, p. 1672–1676. doi:10.1145/3404835.3462989.
35. Guo, T.; Guo, S.; Wang, J. pFedPrompt: Learning Personalized Prompt for Vision-Language Models in Federated Learning. Proceedings of the ACM Web Conference 2023, 2023, pp. 1364–1374.
36. Chen, Y.; Hsu, C.F.; Tsai, C.C.; Hsu, C.H. HPFL: Federated Learning by Fusing Multiple Sensor Modalities with Heterogeneous Privacy Sensitivity Levels. Proceedings of the 1st International Workshop on Methodologies for Multimedia; Association for Computing Machinery: New York, NY, USA, 2022; M4MM '22, p. 5–14. doi:10.1145/3552487.3556438.
37. Zhang, R.; Chi, X.; Liu, G.; Zhang, W.; Du, Y.; Wang, F. Unimodal Training-Multimodal Prediction: Cross-modal Federated Learning with Hierarchical Aggregation. *arXiv preprint arXiv:2303.15486* **2023**.
38. Yu, Q.; Liu, Y.; Wang, Y.; Xu, K.; Liu, J. Multimodal Federated Learning via Contrastive Representation Ensemble. The Eleventh International Conference on Learning Representations, 2023.
39. Lu, W.; Hu, X.; Wang, J.; Xie, X. FedCLIP: Fast Generalization and Personalization for CLIP in Federated Learning. *arXiv preprint arXiv:2302.13485* **2023**.
40. Liang, P.P.; Lyu, Y.; Fan, X.; Mo, S.; Yogatama, D.; Morency, L.P.; Salakhutdinov, R. HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning, 2022. doi:10.48550/ARXIV.2203.01311.
41. Liang, P.P.; Liu, T.; Cai, A.; Muszynski, M.; Ishii, R.; Allen, N.; Auerbach, R.; Brent, D.; Salakhutdinov, R.; Morency, L.P. Learning language and multimodal privacy-preserving markers of mood from mobile data. *arXiv preprint arXiv:2106.13213* **2021**.
42. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
43. Reed, S.; Akata, Z.; Lee, H.; Schiele, B. Learning deep representations of fine-grained visual descriptions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 49–58.
44. Nilsback, M.E.; Zisserman, A. Automated Flower Classification over a Large Number of Classes. Indian Conference on Computer Vision, Graphics and Image Processing, 2008.
45. Bossard, L.; Guillaumin, M.; Van Gool, L. Food-101 – Mining Discriminative Components with Random Forests. European Conference on Computer Vision, 2014.
46. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context, 2015, [arXiv:cs.CV/1405.0312].
47. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* **2014**, *2*, 67–78.
48. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence* **2019**, *42*, 2684–2701.

49. Damen, D.; Doughty, H.; Farinella, G.M.; Furnari, A.; Ma, J.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; Wray, M. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision (IJCV)* **2022**, *130*, 33–55.
50. Nakamura, K.; Yeung, S.; Alahi, A.; Fei-Fei, L. Jointly learning energy expenditures and activities using egocentric multimodal signals. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1868–1877.
51. Banos, O.; Garcia, R.; Saez, A. MHEALTH Dataset. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5TW22.
52. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* **2008**, *42*, 335–359.
53. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
54. Liang, P.P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.Y.; Wu, P.; Lee, M.A.; Zhu, Y.; others. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.
55. Johnson, A.; Bulgarelli, L.; Shen, L.; et al.. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data* **2023**, *10*, 1. doi:10.1038/s41597-022-01899-x.
56. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, e215–e220.
57. Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI Database. http://adni.loni.usc.edu, 2023.
58. Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J. 3d shapenets: A deep representation for volumetric shapes. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1912–1920.
59. Duarte, M.F.; Hu, Y.H. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing* **2004**, *64*, 826–838. doi:10.1016/j.jpdc.2004.03.020.
60. Feng, T.; Bose, D.; Zhang, T.; Hebbar, R.; Ramakrishna, A.; Gupta, R.; Zhang, M.; Avestimehr, S.; Narayanan, S. FedMultimodal: A Benchmark For Multimodal Federated Learning. *arXiv preprint arXiv:2306.09486* **2023**.
61. Guo, W.; Wang, J.; Wang, S. Deep multimodal representation learning: A survey. *IEEE Access* **2019**, *7*, 63373–63394.
62. 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**.
63. Wang, W.; Tran, D.; Feiszli, M. What makes training multi-modal classification networks hard? Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12695–12705.
64. Ma, M.; Ren, J.; Zhao, L.; Testuggine, D.; Peng, X. Are Multimodal Transformers Robust to Missing Modality? Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18177–18186.
65. Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; Peng, X. Smil: Multimodal learning with severely missing modality. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, Vol. 35, pp. 2302–2310.
66. Wu, M.; Goodman, N. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems* **2018**, *31*.
67. Tsai, Y.H.H.; Liang, P.P.; Zadeh, A.; Morency, L.P.; Salakhutdinov, R. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* **2018**.
68. Rieke, N.; Hancox, J.; Li, W.; Milletari, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; others. The future of digital health with federated learning. *NPJ digital medicine* **2020**, *3*, 119.
69. 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* **2022**, *82*, 102585. doi:10.1016/j.media.2022.102585.
70. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.

71. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; others. Learning transferable visual models from natural language supervision. International conference on machine learning. PMLR, 2021, pp. 8748–8763.
72. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. International Conference on Machine Learning. PMLR, 2022, pp. 12888–12900.
73. Tian, Y.; Wan, Y.; Lyu, L.; Yao, D.; Jin, H.; Sun, L. FedBERT: when federated learning meets pre-training. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2022**, *13*, 1–26.
74. Tan, Y.; Long, G.; Ma, J.; Liu, L.; Zhou, T.; Jiang, J. Federated learning from pre-trained models: A contrastive learning approach. *arXiv preprint arXiv:2209.10083* **2022**.
75. Zhao, H.; Du, W.; Li, F.; Li, P.; Liu, G. FedPrompt: Communication-Efficient and Privacy-Preserving Prompt Tuning in Federated Learning. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5. doi:10.1109/ICASSP49357.2023.10095356.
76. Zhuang, W.; Wen, Y.; Zhang, S. Divergence-aware federated self-supervised learning. *arXiv preprint arXiv:2204.04385* **2022**.
77. Saeed, A.; Salim, F.D.; Ozcelebi, T.; Lukkien, J. Federated self-supervised learning of multisensor representations for embedded intelligence. *IEEE Internet of Things Journal* **2020**, *8*, 1030–1040.
78. Jeong, W.; Yoon, J.; Yang, E.; Hwang, S.J. Federated semi-supervised learning with inter-client consistency & disjoint learning. *arXiv preprint arXiv:2006.12097* **2020**.
79. 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. 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 715–724. doi:10.1109/BigData52589.2021.9671374.
80. Long, Z.; Che, L.; Wang, Y.; Ye, M.; Luo, J.; Wu, J.; Xiao, H.; Ma, F. FedSiam: Towards adaptive federated semi-supervised learning. *arXiv preprint arXiv:2012.03292* **2020**.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.