Personalized Federation Learning with Model-Contrastive Learning for Multi-Modal User Modeling in Human-Centric Metaverse

Xiaokang Zhou, *Member, IEEE,* Qiuyue Yang, Xuzhe Zheng, Wei Liang, *Member, IEEE,* Kevin I-Kai Wang, *Member, IEEE,* Jianhua Ma, *Senior Member, IEEE,* Yi Pan, *Senior Member, IEEE* and Qun Jin, *Senior Member, IEEE*

***Abstract*—With the flourish of digital technologies and rapid development of 5G and beyond networks, Metaverse has become an increasingly hotly discussed topic, which offers users with multiple roles for diversified experience interacting with virtual services. How to capture and model users' multi-platform or cross-space data/behaviors become essential to enrich people with more realistic and immersed experience in Metaverse-enabled smart applications over 5G and beyond networks. In this study, we propose a Personalized Federated Learning with Model-Contrastive Learning (PFL-MCL) framework, which may efficiently enhance the communication and interaction in human-centric Metaverse environments by making use of the large-scale, heterogeneous, and multi-modal Metaverse data. Differing from the conventional Federated Learning (FL) architecture, a multi-center aggregation structure to learn multiple global models based on the changes of dynamically updated local model weights, is developed in global, while a hierarchical neural network structure which includes a personalized module and a federated module to tackle both issues on data heterogeneity and model heterogeneity, is designed in local, so as to enhance the performance of PFL with unique characteristics of Metaverse data. In particular, a two-stage iterative clustering algorithm with a more precise initialization is developed to facilitate the personalized global aggregation with dynamically updated multiple aggregation centers. A personalized multi-modal fusion network is constructed to greatly reduce the computational cost and feature dimensions from the high-dimensional heterogeneous inputs for more efficient cross-modal fusion, based on a hierarchical shift-window attention mechanism and a newly designed bridge attention mechanism. A MCL scheme is then incorporated to speed up the model convergence with less communication overload between the local federated module and global model, while an embedding layer which effectively enables the delivered global model to better adapt to the local personality in each client is further integrated. Compared with five baseline methods, experiment and evaluation results based on two different real-world datasets demonstrate the excellent performance of our proposed PFL-MCL model in a fine-grain personalized training strategy, toward more efficient communication and networking among human-centric Metaverse enabled smart applications.**

***Index Terms*—Personalized Federated Learning, Model-Contrastive Learning, Attention Mechanism, Multi-Modal Fusion, Human-Centric, Metaverse**

Xiaokang Zhou is with the Faculty of Data Science, Shiga University, Hikone 522-8522, Japan, and also with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan (e-mail: zhou@biwako.shiga-u.ac.jp).

Qiuyue Yang is with the Changsha Social Laboratory of Artificial Intelligence, Hunan University of Technology and Business, Changsha 410205, China (e-mail: yangqiuyue233@163.com).

Xuzhe Zheng is with the School of Frontier Crossover Studies, Hunan University of Technology and Business, Changsha 410205, China (e-mail: xuzhezheng245@gmail.com).

Wei Liang is with the Xiangjiang Laboratory, Changsha 410205, China, and also the Changsha Social Laboratory of Artificial Intelligence, Hunan University of Technology and Business, Changsha 410205, China (e-mail: weiliang@csu.edu.cn).

Kevin Wang is with the Department of Electrical, Computer and Software Engineering, The University of Auckland, Auckland 1010, New Zealand (e-mail: kevin.wang@auckland.ac.nz).

Jianhua Ma is with the Faculty of Computer and Information Sciences, Hosei University, Chiyoda-ku 102-8160, Japan (e-mail: jianhua@hosei.ac.jp).

Yi Pan is with the Department of Computer Science, Georgia State University, Atlanta, GA 30302 USA (e-mail: yipan@gsu.edu).

Qun Jin is with the Faculty of Human Sciences, Waseda University, Tokorozawa 359-1192, Japan (e-mail: jin@waseda.jp).

I. Introduction

The advancement of digital technologies has enabled a virtualized world of reality which we considered Metaverse [1]. The Metaverse allows its users to own and use their personalized avatar to feel the virtual reality with immersed interaction experience. Through the Metaverse, users can travel, have social interaction, work, and perform many other activities as they do in real life. Virtual reality, augmented reality, and mixed reality technologies allow better and more convenient interfaces to interact with the Metaverse, and also offer a better experience for users with different roles such as participants and creators. However, virtual services offered in Metaverse also brought forward the challenges such as the requirements of high processing capability and communication resource allocation. Every user in the Metaverse is unique and can be characterized by their behaviors in the Metaverse. In order to provide more realistic, intelligent, and immersed experience in Metaverse, records of historical activities in Metaverse will need to be integrated with vast amount of multi-modal user data coming from heterogeneous platforms. Therefore, it is imperative that Metaverse applications are able to make use of the unique heterogeneous data to offer personalized services and experiences.

The development of deep learning allows the possibility to construct the highly accurate modeling based on user's historical activity records in the Metaverse. However, the computation and communication requirements of deep learning and centralized machine learning are also higher. Federated Learning (FL), as a distributed machine learning approach, has achieved a good success on distributed model training and user privacy protection [2]. Among the users, there exist large amounts of imbalanced and non-IID (independent and

identically distributed) data. Traditional FL has the vulnerability of not able to handle non-IID data and hence results in low model quality and accuracy, but higher communication cost. The traditional centralized training also means the model lacks personalization [3], [4]. For example, traditional FL focuses more on the knowledge sharing and extraction across multiple participants in achieving a better global model, different features and characteristics across different users are therefore lost. The modeling of different Metaverse user can be very different and do not obey the IID assumption, and a good global model will not achieve an equal performance on individual users. Accordingly, the version of Personalized Federated Learning (PFL) is emerging and becomes required, which can construct personalized models for each user based on the global model and hence preserve personalized features, especially in human-centric Metaverse services and applications.

In addition to personalized services, human-centric Metaverse data also has some unique characteristics that need to be tailored in order to achieve more accurate modelling and service provision. First, as previously mentioned, Metaverse client data may come from multiple platforms and therefore is highly heterogeneous and multi-modal. Such multi-modal data results in wider differences across local trained models, and hence leads to the issue of potential slow convergence in the overall FL process. Second, it is critical to achieve real-time learning and data processing in order to support good user experience and human-centric Metaverse applications. Considering these two characteristics, it becomes important to make sure that the heterogeneous multi-modal data can be properly incorporated into local models, while fast convergence can be achieved in PFL to support real-time service provision. To accommodate these challenges, we took inspiration on the traditional contrastive learning and apply that on the model level to form the so-called Model-Contrastive Learning (MCL) to facilitate finding commonality across local models, and therefore ensuring fast convergence. The ability of MCL to handle heterogeneous multi-modal data can completement PFL to make use of the unique characteristics of the diverse Metaverse client data and to enable efficient convergence in the overall PFL process.

In this study, we focus on designing and developing a PFL framework, which may enhance the multi-modal user modeling and personalized recommendation in the human-centric Metaverse environments, by making use of the large-scale, heterogeneous, and multi-modal Metaverse data. In particular, we design and propose a PFL with Model-Contrastive Learning (PFL-MCL) model, in order to enhance the human-centric communication and interaction in Metaverse-enabled smart applications. Compared with the conventional FL architecture, we develop a multi-center aggregation structure to learn multiple global models considering the changes of dynamically updated weights in global, and design a hierarchical neural network structure, including a personalized module and a federated module, to handle the data heterogeneity and model heterogeneity issues in local, which can better facilitate the PFL with Metaverse data. A MCL scheme is incorporated to accelerate the model convergence in local training when facing the imbalanced data distribution from multiple modalities, while a two-stage iterative clustering scheme is developed to realize the multi-center global aggregation for PFL in a more efficient way. Main contributions of this paper can be summarized as follows.

i) A PFL framework is newly designed, which includes a multi-center global aggregation structure facilitated by a two-stage iterative clustering scheme and a hierarchical local training structure integrated with a MCL scheme, so as to improve the recommendation accuracy and reduce the communication overhead with accelerated convergence speed in human-centric Metaverse environments.
ii) A multi-center global aggregation mechanism is developed, in which a so-called two-stage iterative clustering algorithm is devised to enhance the aggregation efficiency with a more precise initialization according to the weight changes in each client, and dynamically update the multiple aggregation centers as well as their involved clients for personalized global aggregation reflecting distribution of local personal data.
iii) A personalized multi-modal fusion network is constructed, in which a hierarchical shift-window attention mechanism is improved to fuse the high-dimensional inputs from users' multi-modal data while effectively reduce the feature dimensions, and a so-called bridge attention mechanism is introduced to greatly reduce the computational cost for more efficient cross-modal fusion from the heterogeneous data.
iv) A MCL scheme is developed to enhance a faster convergence but less communication cost between the local client and global model, in which an embedding layer is further involved to make the delivered global model better adapt to the local personality in each client for PFL.

The rest of the paper will be organized as follows. Section II summarizes the state-of-the-art research works related to FL and Metaverse applications. Section III introduces the problem investigated in the study and the basic framework architecture for PFL. Section IV addresses the implementation of our proposed model with the detailed mechanisms and algorithms. Experiment and evaluation results are demonstrated in Section V, and we conclude this study with promising future directions in Section VI.

II. Related Works

In this section, we give a brief survey on PFL, contrastive learning, and Metaverse application, respectively.

A. Personalized Federated Learning

Comparing with traditional FL paradigms, an emerging concept of PFL, has drawn increasing attentions on dealing with data privacy challenges with personalization strategies, especially when facing highly heterogeneous data in varying edge computing environments [5]. Jin et al. [6] addressed a PFL framework integrated with a self-knowledge distillation scheme, which allowed the local model to be initialized by the global model, and transfer the historical knowledge based on the previous personalized model using self-distillation

technology. Zhou et al. [7] introduced a 2-dimensional FL framework to facilitate the personalized human activity recognition when handling the insufficient training data in cyber-physical-social systems. They incorporated the vertical FL and horizontal FL schemes together, which were employed to extract features from heterogeneous data generated across multiple IoT devices, and aggregate the encrypted local models among multiple individual users respectively. Tashakori et al. [8] developed a PFL model for multi-sensory classifications based on a semi-supervised training scheme. They designed a personalized autoencoder for each user from a hyper network in the cloud server, then generated a series of base models which would be delivered to local training according to different user distributions using their own labeled datasets. Li et al. [9] presented a cluster-based PFL scheme, which used a reinforcement learning enhanced clustering algorithm to group user devices with similar preference, and employed the hierarchical transfer learning to improve the model accuracy, in order to balance the accuracy-cost optimization issue in wireless network environments with multiple base stations. Huang et al. [10] constructed a federated dual network, which included an execution network to obtain the ideal model updating, and an evaluation network to generate the personalized local model under the local application scenario. They further developed a so-called personalized update algorithm with an optimal backtracking replacement policy, to improve the accuracy degradation and stability in FL process. To enable the trained global model that could better adapt the data distribution of each individual client, Farnia et al. [11] integrated the optimal transport theory into the FL scheme, which could transfer samples from multiple distributions to a common probability domain based on the combination of the global model and the learned optimal transport maps. Taking advantage of generated adversarial networks, Cao et al. [12] proposed a PFL scheme to deal with the non-IID data, which allowed local models to be built independently in each client, but no need to share the model structure and parameters with other clients. Mills et al. [13] added the multi-task learning scheme into the general iterative FL framework, and used the non-federated private batch normalization layer to realize the personalization, which could improve the individual model accuracy and convergence speed comparing with the traditional federated averaging algorithm. Yu et al. [14] built an online FL framework to support the personalized federated human activity recognition using a semi-supervised strategy. They developed algorithms to calculate the unsupervised gradient under the consistency training proposition, which could improve the concept drift and convergence instability in an unsupervised gradient aggregation process.

B. Contrastive Learning

Currently, as one especial self-supervised representation learning, contrastive learning has shown growing popularity and great promise in richer vector representation, which may collaborate with many other AI-related technologies, including metric learning, knowledge distillation, relational reasoning, etc, for complex network services. Compared with conventional supervised learning schemes, Chen et al. [15] developed a contrastive self-supervised learning algorithm of visual representation, in which a so-called learnable nonlinear transformation was addressed between the representation and contrastive loss, to improve the feature representation quality without needing a specialized architecture. Wu et al. [16] presented a graph contrastive learning network to facilitate the unsupervised cross-domain classification, which leveraged the attraction and repulsion forces for the intra- and inter-domain consistency, enabling the knowledge transferring among different domains for better embedding feature learning. Wang and Qi [17] built a so-called contrastive learning with stronger augmentations framework, in which they designed a distributional loss to enhance the knowledge transfer from weakly augmented views to strongly augmented views, in order to improve the representation of weakly augmented images. Kermiche [18] introduced a contrastive Hebbian feedforward learning scheme for Boltzmann machines, which could be employed to improve the training of deep neural network based on the estimation only based on feedforward computations, local contrastive Hebbian correlations, and local disturbances. Wang et al. [19] integrated the clustering scheme into the contrastive learning framework for human activity recognition, which could select the same-cluster samples from negative pairs based on a newly defined contrastive loss function. Zhu et al. [20] combined the reinforcement learning and contrastive learning together, and addressed a multi-instance reinforcement contrastive learning framework, in which a reinforcement learning-based agent was designed to assist the contrastive learning via better selection of the discriminative feature sets with inherent semantic relationships. Liu et al. [21] constructed a contrastive self-supervised learning framework to deal with graph anomaly detection tasks on attributed networks. They defined a so-called contrastive instance pair which could capture the node's local information as well as its neighboring substructure, in order to improve the learning of representative information between pairs of node-subgraph instances. Zhu et al. [22] addressed a contrastive representation method to enhance a reinforcement learning framework, which considered the correlation among consecutive inputs, and jointly trained the CNN encoder and Transformer through a contrastive learning process, in order to reconstruct features based on context frames. He et al. [23] proposed a graph contrastive learning model, in which a contrastive learning scheme was developed to train the attribute completion and representation learning in an unsupervised heterogeneous framework, aiming to handle the missing attributes and jointly learn the embeddings of nodes and attributes.

C. Metaverse Applications

As an embodied version of Internet, the concept of Metaverse has been widely discussed in both academic and industry fields, as an exceptional multi-dimensional and multi-sensory communication medium in end-edge-cloud environments [24]. Meng et al. [25] presented a co-design framework of sampling, prediction, and communication, aiming to synchronize device trajectories in both real world and its digital world

for Metaverse. They developed a deep reinforcement learning algorithm with expert knowledge to improve the sampling rate and prediction horizon as well. Han et al. [26] addressed a hierarchical data collection framework with a group of IoT-assisted digital twins for Metaverse. They employed an evolutionary game strategy to model the device selection behaviors in both Metaverse and physical world components for the synchronization optimization. Deveci et al. [27] investigated three implementation options for integrating autonomous vehicles in Metaverse, which were evaluated based on a multi-criteria decision-making method using the q-rung orthopair fuzzy sets, and could be used to enhance the personal mobility in terms of vehicle assessment in Metaverse. Jiang et al. [28] discussed a coded distributed computing framework for Metaverse in vehicular services based on blockchain technologies. They defined a reputation metric for reliability evaluation, and applied a game-theoretic method to find a sustainable scheme to improve user experience in vehicle Metaverse. Aiming to investigate the social and educational impact of Metaverse, Wang et al. [29] proposed a theoretical framework with four basic components, to review literatures and synthesize learning practices for education Metaverse ecosystem. Bansal et al. [30] surveyed the state-of-the-art Metaverse applications in healthcare industry, including seven domains as: telemedicine, clinical care, education, mental health, physical fitness, veterinary, and pharmaceuticals, which pointed out technical issues and directions for future development of Metaverse in medical and healthcare-related systems. Ren et al. [31] introduced a so-called quantum collective learning method with a matching game theory to model connected and autonomous vehicles in Metaverse. They defined the spectrum resource allocation problem as a discrete Markov decision process, and developed a quantum-inspired reinforcement learning mechanism to optimize the distributed vehicle selection policy. Shi et al. [32] employed multi-agent reinforcement learning scheme to model the collective intelligence in digital entity, aiming to enhance the immersive environment in Metaverse. They implemented a deep deterministic policy gradient for the domain randomization, which could assist a perception-control modularization for the improvement of generalization performance in multiple unmanned aerial vehicle systems.

III. PFL in Human-Centric Metaverse

In this section, we first discuss basic issues faced in human-centric Metaverse scenarios when building the PFL model. The overall framework of our PFL-MCL with core function modules is then introduced.

A. Application Scenario and Problem Definition

As shown in Fig.1, we introduce the PFL-MCL framework to provide users with more adaptive personalized services in the human-centric Metaverse environment, in which users may generate a large amount of personal data, including browsing and shopping records, game data, and even physiological data (e.g., brain waves and electrocardiograms) detected by wearable sensors. Considering such scenarios, a single client may not only constrain computing resources with limited data access, but also need to face the high requirement of real-time data process with the privacy, security, and heterogeneity issues. Our proposed framework can effectively improve the distributed model training performance on the premise of data privacy protection, while providing each user with more personalized services, which may better adapt to the heterogeneous data in Metaverse-oriented applications. In particular, this framework incorporates the MCL to accelerate the convergence speed in local model training to meet the real-time requirements, and a hierarchical neural network structure is constructed locally to facilitate the personalized design for each client. Given the total model weight indicated as W, the neural network model includes two important components: a personalized module indicated as $W_{\mathrm{P}}(\cdot)$, which is used to extract and fuse multi-modal feature vectors from the heterogeneous input data, and a federated module indicated as $W_{\mathrm{F}}(\cdot)$, which is designed to facilitate the participation of heterogeneous local models for cross-modal fusion in PFL with higher training and communication efficiency based on the contrastive learning strategy.

Generally, the personal data generated by each user in the Metaverse environment is complex and diverse, which may lead to serious data heterogeneity issues between each client, and reduce the training performance when using traditional FL schemes. In such scenario, it is obvious that whenever the client uses its local data to update the model, the personalized local model may then deviate from the aggregated global model. When the local model is later uploaded for global model aggregation, it may impact the global model convergence, resulting in low training efficiency for PFL. Therefore, we improve the FL with a multi-center global aggregation mechanism to better address the heterogeneity by assigning clients to different clusters. Specifically, a two-stage clustering algorithm is developed in global to aggregate similar $W_{\mathrm{F}}(\cdot)$ into the same center in an iterative way. Given K as the the number of centers, it is noted that, two extreme cases, i) when $K = 1$, it is the original federated learning with one aggregated model(i.e., same to the traditional FedAvg), which is not easy to capture the heterogeneous features from different clients, and cannot adapt well to the personalized model training in a specific client; ii) when $K = m$ (m is the number of clients), it becomes a m-center aggregation model, which means every client has its own aggregation model. This is, however, not federated learning and need to be avoided. Therefore, the goal is to find a suitable number for K during the multi-center global aggregation process for more efficient PFL.

During the local training in the PFL-MCL, both the loss function of the local model and the distance between the local model and the global model, need to be considered. Thus the MCL scheme is involved, which can not only speed up the model convergence, but also ensure the local model is not too far away from the global model. The optimization objective is to minimize the loss function ℓ_{local} in terms of the m-th local model and can be formulated as follows.

$$\min_{W_m^t} \ell_s(D_m, M_m^t, W_m^t) + \ell_c(D_m, W_{\mathrm{F}m}^t, W_{\mathrm{F}m}^{t-1}, W_{\mathrm{G}k}^t) \quad (1)$$

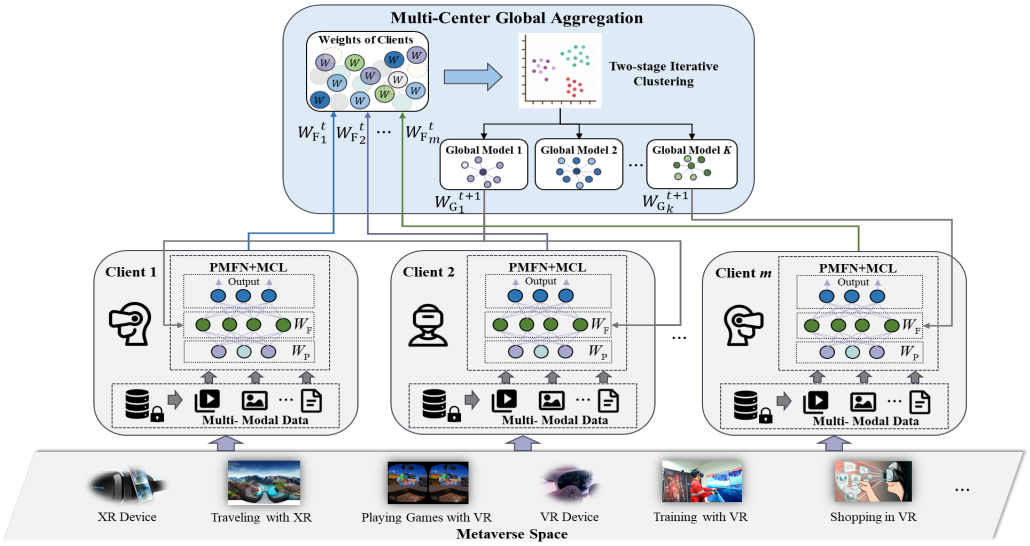

Fig. 1. Multi-Modal Modeling Based on PFL-MCL in Human-Centric Metaverse

where $\ell_s(\cdot)$ is the supervised loss function of the local model, $\ell_c(\cdot)$ is the MCL loss function. $D_m = \{x_m, y_m\}$ is the private data set of m-th client, x_m and y_m denote the input sample and the corresponding label respectively. M_m^t denotes the local neural network model of the m-th client in the t-th round, $M_m^t : x_m \rightarrow y_m$, which can be used to predict user behavior and provide personalized recommendations for users in the Metaverse environment. W_m^t denotes the total model weight of the m-th client in the t-th round. $W_{\mathrm{F}m}^{t-1}$ and $W_{\mathrm{F}m}^t$ denote the weights of federated module of the m-th client in the (t-1)-th and t-th rounds respectively. $W_{\mathrm{G}k}^t$ denotes the global weight of aggregated federated modules from the k-th global aggregation center in the t-th round.

Accordingly, the optimization problem of our PFL-MCL in global can be formulated as follows.

$$\min_{\{W_i^t\},\{r_{ik}\},\{W_{\mathrm{G}k}^t\}} \sum_{i=1}^{m} (\alpha_i \ell_{\text{local}} \left(M_i^t, D_i, W_i^t, W_{\mathrm{F}i}^t, W_{\mathrm{F}i}^{t-1}, W_{\mathrm{G}k}^t\right) + \frac{\lambda}{m} \sum_{k=1}^{K} \sum_{i=1}^{m} r_{ik} Dist\left(W_{\mathrm{F}i}^t, W_{\mathrm{G}k}^t\right)) \tag{2}$$

where $\alpha_i = \frac{|D_i|}{\sum_j |D_j|}$ indicates an importance weight measured by the number of samples in each client. r_{ik} indicates a judgement function to determine whether the i-th client belongs to the k-th center or not. λ is a coefficient to control the trade-off between ℓ_{local} and the distance indicated as $Dist(\cdot)$.

It is noted that the second term in the Eq.(2) aims to minimize the distance between each federated module in local and its corresponding nearest global center, thus can be optimized by minimizing the intra-cluster distance and described as follows.

$$\min_{\{r_{ik}\},\{W_{\mathrm{G}k}^t\}} \frac{1}{m} \sum_{k=1}^{K} \sum_{i=1}^{m} r_{ik} Dist(W_{\mathrm{F}i}^t, W_{\mathrm{G}k}^t) \tag{3}$$

B. Overall Framework

As shown in Fig.2, a two-layer structured framework of PFL-MCL is constructed, which includes a hierarchical neural network structure in local, and a multi-center aggregation structure in global. Specifically, the training of local model consists of two essential components, namely, a personalized module and a federated module, aiming to handle the heterogeneous multi-modal data fusion in a personalized manner, while reduce communication overhead with a higher training efficiency. The personalized module is designed to refine the cross-modal fusion with lower computational complexity based on the idea of bridge attention, and the federated module is devised to enhance the personalized local training based on a developed MCL scheme, making it increasingly closer to the global model with accelerated model convergence speed. A multi-center global aggregation mechanism is then developed in global based on a two-stage iterative clustering scheme, in order to improve the aggregation efficiency while alleviate the model deviation issue caused by personal data heterogeneity in PFL.

Basically, the multi-modal data will first be input into the personalized module, and the low-dimensional feature representations are extracted and generated using different encoders through a hierarchical shift-window attention network. A so-called bridge attention structure is further incorporated for more efficient cross-modal fusion from the heterogeneous data. Then, in the federated module, a MCL scheme is developed for the convergence optimization in FL, which allows the federated module to get closer to the global model, but far away from that of the previous round in an accelerated way. An embedding layer is newly designed and added to further enhance the personality in local model training, which enables the delivered global model to adapt to local data features in a more personalized way. Furthermore, a two-stage iterative clustering algorithm is implemented in global to realize the multi-center global aggregation, based on which a suitable number of K centers are dynamically determined and

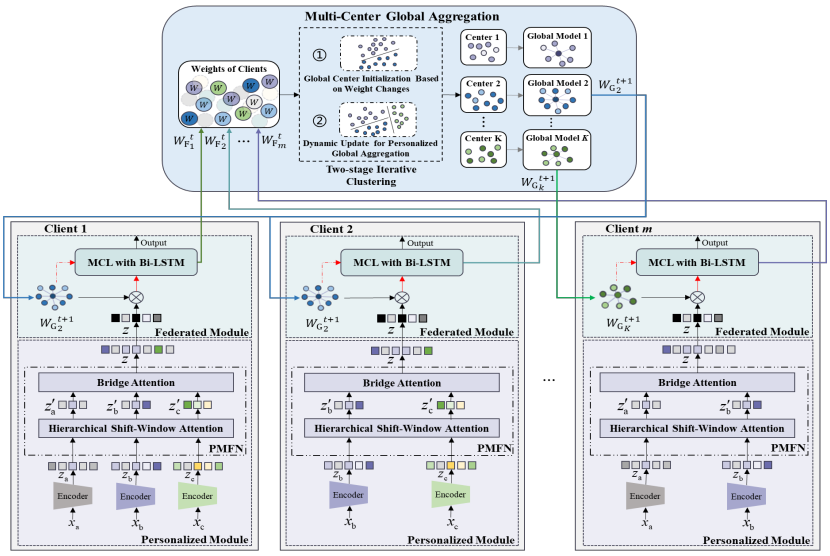

Fig. 2. Overall Architecture of PFL-MCL

initialized considering the changes of weights in local models in the first stage, while the uploaded weights of each client can be assigned to the closest center for more efficient global aggregation with less iterations in the second stage.

With large quantities of heterogeneous data coming from multiple modalities, the computational loads are also expected to be high. Therefore, we design a hierarchical shift-window attention structure, which divides the large number of neurons into a certain number of small windows, then conducts the attention calculation within each small window. It is noted that each divided window is not isolated, but has a certain overlap. This design can help obtain the better time-series feature representation while reduce the dimension with less computational cost. Inspired by the "Attention Bottleneck" in [33], we further develop and construct a bridge attention structure, which incorporates a set of "bridge" neurons between each modality, so as to extract features of adjacent modalities in a more aggregated way based on the restriction of attention flow. Compared with "Attention Bottleneck", our bridge attention mechanism has less computation but higher efficiency, because only the "bridge" neurons need to be involved in the final cross-modal fusion. Moreover, to speed up the convergence in local model training, the MCL is introduced, which is employed to reduce the distance between the local and global models, and increase the distance between the local model and that of the previous round. Compared with [34], we leverage an embedding layer to better adapt the delivered global model to the local personality, which can not only control the deviation in local, but also enhance the real-time performance in PFL. In addition, to maximize the benefit of MCL, a multi-center aggregation structure is adopted in global, which may also better adapt to different clients/users with heterogeneous data from multiple modalities. Compared with [35], a two-stage iterative clustering algorithm with more precise cluster initialization is developed in our PFL-MCL, so as to reduce the communication overhead and improve the robustness of the overall framework with accelerated convergence speed in PFL.

IV. Modeling and Implementation of PFL-MCL

In this section, we mainly discuss the detailed implementation of our proposed PFL-MCL model, especially including the heterogeneous multi-modal fusion, local model training based on MCL, and multi-center global aggregation.

A. Bridge Attention Based Multi-Modal Fusion

As we discussed above, the local training model is composed of a personalized module and a federated module. In the personalized module, to deal with the heterogeneous data in each client, a personalized multi-modal fusion network is designed and constructed, including a hierarchical shift-window attention structure and a so-called bridge attention structure, so as to facilitate the heterogeneous data fusion from multiple modalities in a more efficient way. Specifically, given a client m, the multi-modal data may include the audio, image, and text data, which can be recorded as x_{a}, x_{b}, and x_{c}. To better integrate and fuse these multi-modal data, different encoders will first be used to map the data into low-dimensional features, e.g., using CNN, LSTM, and Transformer, for feature extraction respectively. Then the feature vectors corresponding to x_{a}, x_{b}, x_{c} can be converted into the token sequences, which can be described as follows.

$$z = [z_{\mathrm{a}}||z_{\mathrm{b}}||z_{\mathrm{c}}] \tag{4}$$

$$z_{\mathrm{a}} = f(x_{\mathrm{a}}, E_{\mathrm{a}}), z_{\mathrm{b}} = f(x_{\mathrm{b}}, E_{\mathrm{b}}), z_{\mathrm{c}} = f(x_{\mathrm{c}}, E_{\mathrm{c}}) \tag{5}$$

where $[\cdot||\cdot||\cdot]$ denotes the concatenation of the tokens for each modality. $f(\cdot)$ indicates a mapping relationship. E_{a}, E_{b}, and E_{c} are the encoders of the corresponding modalities respectively.

It is noted that if the token sequence z is directly fused using the traditional attention mechanism, all the neurons need to be considered for each calculation on one neuron, which may lead to a huge computing consumption. Therefore, we design

a hierarchical shift-window attention mechanism to reduce the dimension of the token sequence, which is similar to the downsampling process in CNN. In details, the shift-window mechanism divides the input token sequence into a series of small windows with n neurons in each independent modality according to their corresponding time steps, and the attention calculation will then be performed within each window. Since the divided windows are not independent, as shown in Fig.3, the overlaps existing among them may help increase the interactions between windows, ensuring the information related to some certain time steps can simultaneously appear in multiple windows. In addition, as the hierarchical layer increases, the overlapping parts will gradually decrease, thus can effectively prevent redundancy in feature extraction for the over-fitting issue. The shift-window attention of a specific layer l can be described as follows.

$$z^{l+1} = SWA^l(z^l; \beta) \tag{6}$$

where $SWA^l(\cdot)$ indicates the shift-window attention function of the l-th layer, and $z^{l+1} = [z_{\mathrm{a}}^{l+1}||z_{\mathrm{b}}^{l+1}||z_{\mathrm{c}}^{l+1}]$ is a low-dimensional token sequence. $\beta = \{\beta_{\mathrm{a}}, \beta_{\mathrm{b}}, \beta_{\mathrm{c}}\}$ denotes the independent parameters of each modality.

To avoid repeated calculations and reduce the square complexity of attention mechanism, inspired by [33], we design a bridge attention mechanism, introducing a number of "bridge" neurons between each modality, to better capture the time-series features based on the restriction of attention flow between two adjacent modalities. Additionally, to further reduce the computational complexity in the fusion process, the number of "bridge" neurons B needs to be set as much smaller than the number of neurons N in each modality (i.e., $B \ll N_{\mathrm{a}}$, $B \ll N_{\mathrm{b}}$, $B \ll N_{\mathrm{c}}$). Given the output of the hierarchical shift-window attention network as $z' = [z'_{\mathrm{a}}||z'_{\mathrm{b}}||z'_{\mathrm{c}}]$, the input sequence of bridge attention network with two "bridges" z_1 and z_2 can be described as follows.

$$z' = [z'_{\mathrm{a}} \| z_1 \| z'_{\mathrm{b}} \| z_2 \| z'_{\mathrm{c}}] \tag{7}$$

Since the "bridge attention" is designed to obtain cross-modal information from multi-modal tokens, the cross-modal calculation can be expressed as follows.

$$z^{l+1} = \text{ Cross\_Attention } \left(z'^l; \theta\right) \tag{8}$$

where $\theta = \{\theta_{\mathrm{a}}, \theta_{\mathrm{b}}, \theta_{\mathrm{c}}\}$ is the independent parameters of each modality.

Furthermore, fusion of neurons from different modalities (i.e., z'_{a}, z'_{b}, and z'_{b}) based on the "insert" of z_1 and z_2, can be refined as follows.

$$\begin{aligned}
z_{1,\mathrm{a}}^{l+1} &= \text{Attention}\left(\left[\mathrm{z}'^l_{\mathrm{a}} \| \mathrm{z}^l_1\right]; \theta_{\mathrm{a}}\right) \\
z_{1,\mathrm{b}}^{l+1} &= \text{Attention}\left(\left[\mathrm{z}'^l_{\mathrm{b}} \| \mathrm{z}^l_1\right]; \theta_{\mathrm{b}}\right) \\
z_{2,\mathrm{b}}^{l+1} &= \text{Attention}\left(\left[\mathrm{z}'^l_{\mathrm{b}} \| \mathrm{z}^l_2\right]; \theta_{\mathrm{b}}\right) \\
z_{2,\mathrm{c}}^{l+1} &= \text{Attention}\left(\left[\mathrm{z}'^l_{\mathrm{c}} \| \mathrm{z}^l_2\right]; \theta_{\mathrm{c}}\right)
\end{aligned} \tag{9}$$

where z_{a} and z_{b} exchange information through z_1 to obtain $z_{1,\mathrm{a}}^{l+1}$ and $z_{1,\mathrm{b}}^{l+1}$ respectively, z_{b} and z_{c} exchange information through z_2 to obtain $z_{2,\mathrm{b}}^{l+1}$ and $z_{2,\mathrm{c}}^{l+1}$ respectively.

Then the fused neurons related to the same "bridge" can be merged together, which are described as follows.

$$\begin{aligned}
z_1^{l+1} &= z_{1,\mathrm{a}}^{l+1} \odot z_{1,\mathrm{b}}^{l+1} \\
z_2^{l+1} &= z_{2,\mathrm{b}}^{l+1} \odot z_{2,\mathrm{c}}^{l+1}
\end{aligned} \tag{10}$$

where $\odot$ is the Hadamard product.

Accordingly, the output of the bridge attention network, which may represent features fused from different modalities in a more precise way, based on the $z_1^{(l+1)}$ and $z_2^{(l+1)}$ in the $l+1$ layer, can be described as follows.

$$z_{1,2}^{l+2} = \text{Attention}(z_1^{l+1}, z_2^{l+1}) \tag{11}$$

Differing from [33], which directly input the token sequence into a "bottleneck" structure and needs to repetitively conduct the attention calculation from all the fused neurons, our multi-modal fusion network first incorporates a shift-window attention structure to reduce the dimension of multi-modal tokens, so as to alleviate the computational complexity in terms of the attention calculation during the the further fusion process. Then the idea of bridge attention is involved to refine the cross-modal fusion from multiple modalities, where only the "bridge" neurons need to be considered to generate the final output of the fused features. Finally, a fully connected layer is employed to obtain the feature vectors with unified dimensions, as the normalized input to the next federated module.

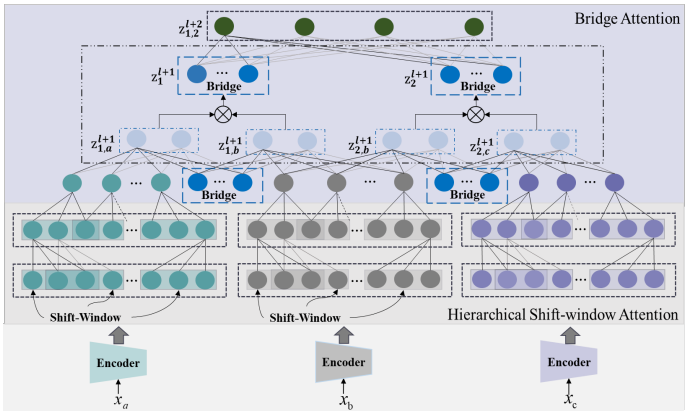

Fig. 3. Conceptual Image of Hierarchical Shift-Window Attention and Bridge Attention Mechanisms

B. MCL Enhanced Local Training

Basically, the Bi-LSTM structure, which is an extension of the traditional LSTM network, allowing information to flow forward and backward in the network, is adopted to enhance the sequential context-aware feature extraction in the federated module.

On this basis, we develop a MCL scheme to enhance the personalized local training, while optimizing the convergence in FL. Compared with traditional contrastive learning, we conduct it in the model level, making the federated module in local continuously to become closer to the global model in the corresponding aggregation center, but further away from that of the previous round. In addition, differing from [34], an

embedding layer is newly added, which allows the delivered global model to adapt to local data features, then speed up the local training based on MCL. This design can not only enhance the convergence speed, but also control the deviation between the local federated module and the global model, so as to ensure the real-time performance in each client.

As shown in Fig.4, in the t-h round of communication, after receiving $W_{\mathrm{G}k}^t$ from the global aggregation model, the client m first makes $W_{\mathrm{G}k}^t$ adapt to the local data feature through the embedding layer. The calculation in the embedding layer can be described as follows.

$$W_{\mathrm{F}m}^t = W_{\mathrm{G}k}^t \odot FC(z_{1,2}^{l+2}) \tag{12}$$

where $W_{\mathrm{F}m}^t$ denotes the weight of the local federated module of client m in the t-th round. $FC(\cdot)$ denotes the function of the fully connected layer.

It is noted that when using the local data to update the model, two kinds of losses may need to be considered. The first one is the typical loss in supervised learning (e.g., binary cross-entropy loss), which is denoted as ℓ_s. While the second one is our MCL loss, which is denoted as ℓ_c, and can be specified as follows.

$$\ell_c = -\log\left(\frac{\exp(sim(W_{\mathrm{F}m}^t, W_{\mathrm{G}k}^t)/(\tau\Delta t))}{\frac{\exp(sim(W_{\mathrm{F}m}^t, W_{\mathrm{G}k}^t))}{\tau\Delta t} + \frac{\exp(sim(W_{\mathrm{F}m}^t, W_{\mathrm{F}m}^{t-1}))}{\tau\Delta t}}\right) \tag{13}$$

where Δt denotes the time difference between the t-th model update and the (t-1)-th model update. τ denotes a so-called temperature parameter. The numerator is the positive sample pair $(W_{\mathrm{F}m}^t, W_{\mathrm{G}k}^t)$, and the denominator is all sample pairs, including positive sample pairs and negative sample pairs $(W_{\mathrm{F}m}^t, W_{\mathrm{F}m}^{t-1})$. $W_{\mathrm{F}m}^{t-1}$ denotes the federated module in the (t-1)-th round. $sim(\cdot)$ denotes the function of similarity calculation.

The specific calculation of similarity can be formulated as follows.

$$sim(W_{\mathrm{F}m}^t, W_{\mathrm{G}k}^t) = \frac{1}{n}\sum_{j=1}^{n} \frac{A(W_{\mathrm{F}mj}^t) \times A(W_{\mathrm{G}kj}^t)}{\sqrt{|A(W_{\mathrm{F}mj}^t)|} \times \sqrt{|A(W_{\mathrm{G}kj}^t)|}} \tag{14}$$

where $A(\cdot)$ is a function that expands the matrix into a one-dimensional vector. n is the number of weight matrices of the federated module.

Accordingly, the total local loss function of the m-th client can be expressed as follows.

$$\ell_{\text{local}} = \ell_s\left(M_m^t, D_m, W_m^t\right) + \mu\ell_c\left(W_{\mathrm{F}m}^t, W_{\mathrm{F}m}^{t-1}, W_{\mathrm{G}k}^t\right) \tag{15}$$

where μ is a hyperparameter that controls the MCL loss.

The overall algorithm for the local model training is shown in Algorithm 1. In the t-th round of FL, after the client receives the delivered $W_{\mathrm{G}k}^t$, it will map $W_{\mathrm{G}k}^t$ to $W_{\mathrm{F}m}^t$ through the embedding layer. Then in each epoch, the local data set D_m is used to update the local model using the stochastic gradient descent, and $W_{\mathrm{F}m}^t$ for the federated module will be updated through ℓ_{local}, and uploaded for the next round of FL.

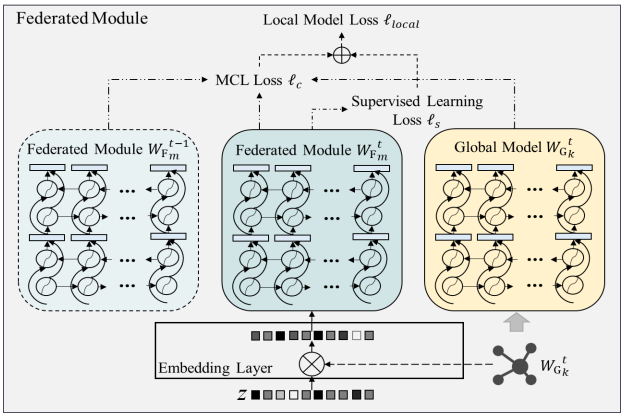

Fig. 4. MCL Scheme for PFL

Algorithm 1 Personalized Local Model Training

Input: Dataset D_m
Output: The trained model M_m^T after T rounds

Initialize the number of communication rounds T
Initialize the number of clients m, the number of local epochs E
Initialize temperature τ, learning rate η, and hyper-parameter μ
$LocalTraining(D_m, W_{\mathrm{G}k}^t)$:
for $t = 1, 2, \ldots, T$ **do**
  Receive $W_{\mathrm{G}k}^t$ from global
  $W_{\mathrm{F}m}^t \leftarrow W_{\mathrm{G}k}^t$
  for epoch $e = 1, 2, ..., E$ **do**
    for each batch $b = \{x_m, y_m\}$ of D_m **do**
      Calculate the ℓ_s
      Calculate the ℓ_c based on Eq. (15)
      $W_{\mathrm{P}m}^t \leftarrow W_{\mathrm{P}m}^t - \eta\nabla\ell_{local}$
      $W_{\mathrm{F}m}^t \leftarrow W_{\mathrm{F}m}^t - \eta\nabla\ell_{local}$
      $M_m^t \leftarrow W_{\mathrm{P}m}^t + W_{\mathrm{F}m}^t$
    end for
    Upload $W_{\mathrm{F}m}^t$ to global
  end for
end for
Return M_m^T

C. Two-stage Clustering Based Multi-center Global Aggregation

Usually, a model trained on the entire dataset may achieve better feature extraction than that trained on a skewed subset, but when the client faces the very severe data heterogeneity issue, it may lose its advantage in the traditional FL. As considered in [35], the K-Means algorithm was applied to cluster the data from different users and form the multi-center FL. However, if the initial number of clusters is uncertain, too many or too few initially set centers may lead to more iterations to converge to an optimized result.

Therefore, we propose a two-stage iterative clustering algorithm to realize a dynamic multi-center generation process for PFL. Specifically, in the first stage, the number of centers will be initialized based on the changes in terms of the local model

weight, reflecting the distribution of local personal data. Then in the second stage, the uploaded weights of each client will be assigned to the closest center for global aggregation during the following iteration process.

In particular, in the first round of communication, the global model will initialize $W_{\mathrm{G}k}^{1}$ and send it to all clients, the number of centers in global is $K = 1$. After receiving $W_{\mathrm{G}k}^{1}$, for example, the m-th client will use its local personal data to update $W_{\mathrm{F}m}^{1}$ based on $W_{\mathrm{G}k}^{1}$, and upload it back to the global. Then the corresponding weight change between the m-th client and the k-th global center can be calculated as follows.

$$\Delta W_m = \left| W_{\mathrm{G}k}^{1} - W_{\mathrm{F}m}^{1} \right| \tag{16}$$

The cosine distance is used to calculate the distance among the weight changes of each client, which can be formulated as follows.

$$d_{i,j} = \frac{\Delta W_i, \Delta W_j}{\|\Delta W_i\| \, \|\Delta W_j\|} (i \neq j; i, j = 1, 2, \cdots, m) \tag{17}$$

The first stage clustering is then performed to find the suitable K centers, which is similar to the density-based clustering algorithm. Given m clients recorded as object set $S = \{W_{\mathrm{F}1}, W_{\mathrm{F}2}, \ldots, W_{\mathrm{F}m}\}$, First, an object $W_{\mathrm{F}i}$ is randomly selected from S and determine whether it belongs to an existing cluster. If yes, re-select an object from S. Otherwise, if $W_{\mathrm{F}i}$ is identified as the core object point, calculate the distance among it and other objects (e.g., $W_{\mathrm{F}j}$). If the distance is less than the specific threshold and $W_{\mathrm{F}j}$ does not belong to any existing cluster, put $W_{\mathrm{F}j}$ and $W_{\mathrm{F}i}$ into a new cluster. Repeat the above process until all the objects in S are assigned to a certain cluster. Following this way, we initialize a suitable number (i.e., K) of multiple centers, and the multi-center global aggregation can be conducted within each center.

As for the second stage clustering in the following communication rounds, the client will send the updated $W_{\mathrm{F}m}^{t}$ to the global, then the distance between $W_{\mathrm{F}m}^{t}$ and $W_{\mathrm{G}k}^{t}$ in the existing K centers can be calculated. Consequentially, the center with the closest distance will be selected to join, which can be described as follows.

$$r_{mk} = \begin{cases} 1, \ if \ k = \arg\min_j sim(W_{\mathrm{F}m}^{t}, W_{\mathrm{G}j}^{t}) \\ 0, \ otherwise \end{cases} \tag{18}$$

where r_{mk} is defined to determine whether $W_{\mathrm{F}m}^{t}$ can be assigned to the k-th center for global aggregation. $sim(\cdot)$ is a similarity calculation function.

Based on these, the global aggregation will perform the weighted summation within all the K centers respectively, which can be formulated as follows.

$$W_{\mathrm{G}k}^{t+1} = \frac{1}{\sum_{i=1}^{m} r_{ik}} \sum_{i=1}^{m} r_{ik} W_{\mathrm{F}i}^{t} \tag{19}$$

Finally, the global model aggregated in each center can be delivered to the corresponding clients, and the clients will continue to update their own local models, the process of which will be repeated until the convergence is reached in every client.

The specific multi-center global aggregation mechanism is shown in Algorithm 2. The first stage is to initialize the suitable number of global aggregation centers, based on the calculation of dynamic weight changes of local models. The second stage is to measure the distance among uploaded local weights and existing global centers, where the center with the closest distance will be selected to conduct the aggregation. The second stage may be repeated until all the local model training converge based on the delivered global model in each round.

Algorithm 2 Multi-Center Global Aggregation for PFL

Input: Local $W_{\mathrm{F}m}^{t}$ from each client
Output: The global weight $W_{\mathrm{G}k}^{T}$ after T rounds

Initialize the number of communication rounds T
Initialize the number of clients m
Initialize $W_{\mathrm{G}k}^{1}$
for $t = 1, 2, 3, \ldots, T$ **do**
  Deliver $W_{\mathrm{G}k}^{t}$ to each client
  Update $W_{\mathrm{F}m}^{t}$: $W_{\mathrm{F}m}^{t} \leftarrow LocalTraining(D_m, W_{\mathrm{G}k}^{t})$
  if $t = 1$ **then**
    Calculate ΔW_m based on Eq. (16)
    Calculate d_{ij} based on Eq. (17)
    Generate K centers using $d_{i,j}$
    Average aggregation within the center: $W_{\mathrm{G}k}^{t+1} \leftarrow \frac{1}{n} \sum W_{\mathrm{F}m}^{t}$
  end if
  if $t \neq 1$ **then**
    Select the optimal center based on Eq. (18)
    Update $W_{\mathrm{G}k}^{t+1}$ based on Eq. (19)
  end if
end for
Return $W_{\mathrm{G}k}^{T}$

V. EXPERIMENT AND ANALYSIS

In this section, the experimental setup and used dataset will be presented first, followed by the performance evaluations against five baseline methods in terms of stability, efficiency, and accuracy.

A. Experiment Design

The experiments are conducted with two real-world datasets focusing on user behavior prediction for both the offline and online tests. The Amazon Product Reviews (APR) dataset that contains 847,733 interactions involving 70,679 users and 24,915 items is applied for user prediction offline experiment [36]. The APR dataset presents a comprehensive compilation of reviews, product metadata, and inter-product connections within the Amazon platform. It provides detailed information about products, including description, category, price, and brand specifications. Additionally, it features photos showing products that are "also viewed" or "also bought" together. The dataset also encompasses essential review details such as star ratings, review texts, and helpfulness votes, indicating how many users found a review beneficial. For the validation of

our model within the Metaverse environment, we partitioned user behavior features, reserving "Buy", and "Cart" actions in the physical realm, while separating "Click" interactions in the Metaverse space, to ensure they do not overlap with the physical space. The "Buy" and "Cart" actions are considered in the physical space because they represent actual purchase and shopping cart interactions through a real-world interface or device. In contrast, the "Click" action represents interactions such as selecting items and browsing menus that are performed only in the virtual environment.

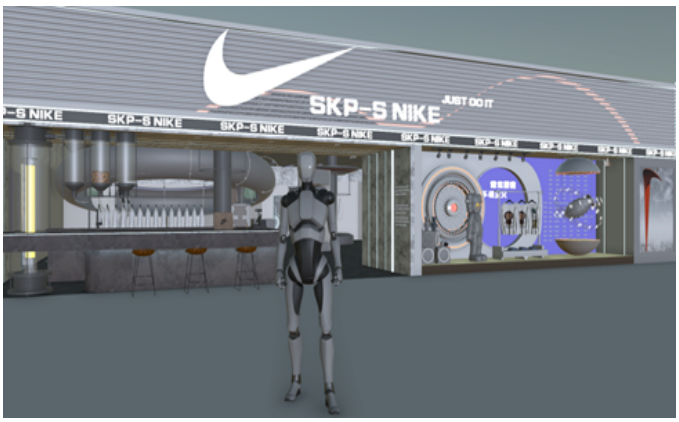

Fig. 5. Image of VSM Metaverse Application

In addition to the APR dataset, we also collected data from the online Virtual Shopping Mall (VSM) in our Metaverse Innovation Laboratory. The data are collected by Unity 3D virtual reality engine and the physiological data acquisition devices, e.g., eye tracking, Electroencephalogram (EEG), Electromyogram (EMG), Electrodermal activity (EDA). Specifically, 154 users from age of 19 to 45 participate in shopping on the VSM platform. All real time behavioral data including user VIP level, shopping records, staying behaviors, trigger times of shopping rewards, and average online time are recorded in the Metaverse space and the physiological data acquisition devices, which are used for the user behavior prediction analysis in this paper.

Both the offline and online datasets used in the article are split into the training set (60% of the total data) and test set (40% of the total data). All experiments were conducted in the 12th Gen Intel (R) Core (TM) i7-12700H 2.70 GHz CPU, 16GB RAM, NVidia GeForce GTX 3060 Ti GPU, python3.9, CentOS environment. To verify the effectiveness of the proposed method, five baseline methods listed below are chosen for comparative analysis. All the methods are performed on both the APR and VSM dataset.

i) NonFed is a conventional centralized machine learning approach that does not involve federated learning. A basic attention mechanism, i.e., the Transfomer model is operated on a single central server in this study, where all data is collected and used for model training.
ii) Multimodal Bottleneck Transformer (MBT) [33] is a specialized model architecture designed for processing multimodal data (e.g., text, images, audio). It utilizes a Bottleneck Transformer structure to efficiently handle the fusion of these diverse data modalities.
iii) Federated Averaging (FedAvg) [37] is a method used to address the model aggregation problem in federated learning. It trains local models on multiple devices or users and then averaging the weights of these models to obtain a global model. This global model exhibits a certain degree of generalization across all participating parties' data.
iV) Moon [34] is a simple and effective federated learning framework that utilizes the similarity between model representations to correct the local training of individual parties, i.e., conducting contrastive learning at the model-level.
V) FeSEM [35] is a multi-center aggregation mechanism to cluster clients using their models' parameters. It learns multiple global models from data as the cluster centers, and simultaneously derives the optimal matching between users and centers. This process is formulated as an optimization problem that can be efficiently solved by a stochastic expectation maximization algorithm.

In the offline comparison test, three general prediction metrics, Precision, Recall and F-Measure are applied to evaluate the performance of all the methods. For a candidate user u with M candidate items, suppose G_u is the ground truth set, which contains the items that are actually relevant to the user u. $\mathcal{P}_u(|\mathcal{P}_u| = M)$ is the recalled set that stands for the set of items recommended by the system for user u. The metrics can be defined as follows.

$$\text{Precision@}M(u) = \frac{|\mathcal{P}_u \cap \mathcal{G}_u|}{\mathcal{P}_u} \tag{20}$$

$$\text{Recall@}M(u) = \frac{|\mathcal{P}_u \cap \mathcal{G}_u|}{\mathcal{G}_u} \tag{21}$$

$$\text{F1-Measure@}M(u) = \frac{2 * \text{Precision@}M(u) * \text{Recall@}M(u)}{\text{Precision@}M(u) + \text{Recall@}M(u)} \tag{22}$$

To evaluate the proposed model in online metaverse environments, the online shopping sbehavior on the VSM platform is examined in the study. The widely used metric Click-Through Rate (CTR) is utilized to measure the performance as follows.

$$\text{CTR@M(u)} = \frac{\text{\# of cart or \# of buy}}{\text{\# of impressions}} \tag{23}$$

CTR measures the ratio of users who click on a specific item (e.g., adding to cart or buying) to the number of times it was recommended and presented to them (i.e., the number of impressions). It is used to assess how well the model performs in terms of generating relevant recommendations that lead to user engagement (clicks). A higher CTR indicates that a larger proportion of users found the recommendations engaging and clicked on them, which is generally considered a positive outcome for the recommendation model.

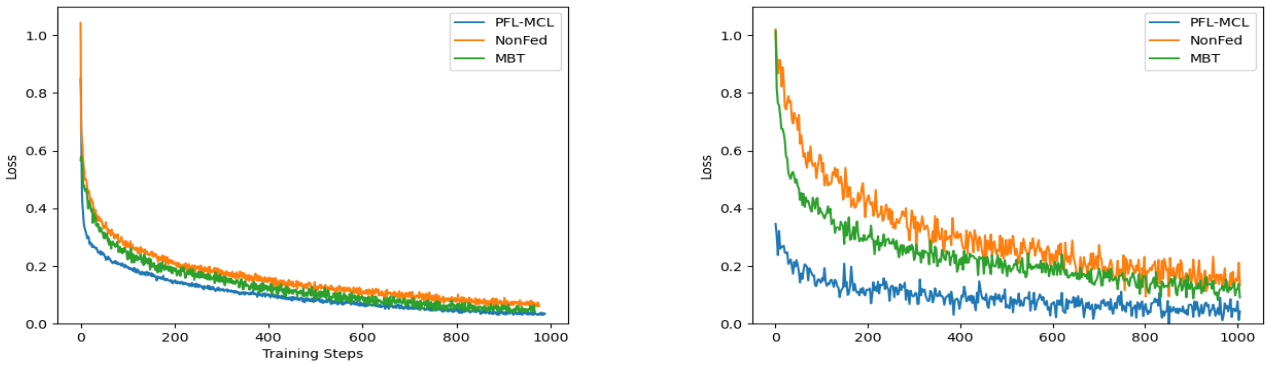

Fig. 6. Local Model Training Performance Evaluation on (a) APR and (b) VSM datasets.

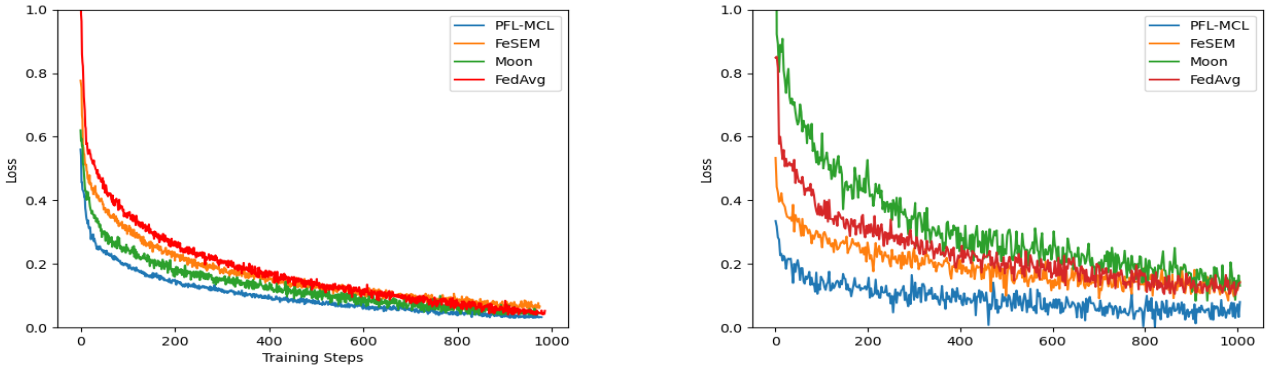

Fig. 7. Federated Learning Training Performance Evaluation on (a) APR and (b) VSM datasets.

B. Evaluation on Local Model Training Performance

To verify the efficiency of the proposed PFL-MCL method, the local model training performance is investigated at first by comparing with NonFed and MBT methods. To obtain a comprehensive training performance evaluation result, both the APR and VSM datasets are used in the evaluation. The local training loss evaluation results are shown in Figs.6(a) and Figs.6 (b) for two datasets respectively.

As shown in Fig. 6, the overall loss consistently decreases with increasing training steps across all methods, suggesting that each approach undergoes a positive learning process on the datasets. For both the VSM and APR datasets, approximately 1000 training iterations are required to reach the designated MCL loss threshold (i.e., 0.05 in the test). Notably, the result of the VSM dataset exhibits more pronounced loss fluctuations compared to the result of the APR dataset. Generally, the PFL-MCL method consistently achieves better results with low loss in both scenarios. In contrast, the other baseline models demonstrate lower learning efficiency observing from the loss curves.

In addition, the local training performance is further analysed by comparing the recommendation effectiveness in the VSM platform. Two widely-used metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are employed here to quantify how well the recommendation system's predictions align with the actual preferences or ratings provided by users. The specific formula is as follow.

$$\mathrm{RMSE}(x, h_i) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h_i(x_i) - y_i\right)^2} \tag{24}$$

$$\mathrm{MAE}(x, h_i) = \frac{1}{m}\sum_{i=1}^{m}\left|h_i(x_i) - y_i\right| \tag{25}$$

where x_i is a user-item pair representing the user and the item they interacted with in a recommendation system; h_i is the prediction made by the recommendation system for the interaction of user x_i with the item; y_i represents the ground truth and m is the total number of data points or instances being considered in the evaluation.

As shown in Table I, the results in the table indicate that all methods' RMSE are more sensitive compared to MAE. This is primarily due to the fact that RMSE gives higher weighting to larger deviations or outliers in the dataset. Overall, PFL-MCL outperforms the other two methods in both RMSE and MAE

TABLE I
PERFORMANCE COMPARISONS FOR DIFFERENT METHODS IN RMSE AND MAE

Method	RMSE	MAE
NonFed	5.2	3.1
MBT	4.8	2.9
PFL-MCL	3.5	2.1

metrics. Additionally, MBT, owing to its capability of fusion of these diverse data modalities in VSM scenario (e.g., eye tracking traces, EGG, EDA, shopping behaviors), exhibits an improvement in error rates to some extent, demonstrating better performance compared to NonFed, which is not designed to handle multi-modal data.

C. Evaluation on Federated Learning Performance

In this section, we further investigate the performance of the PFL-MCL model during the federated learning training process. Following the methodologies addressed in the paper, we evaluate both the model's federated learning aggregation performance and clustering effectiveness. In order to evaluate and compare the federated learning process, the FeSEM, Moon, and FedAvg are selected for evaluation. For the clustering analysis, we compared the results with FeSEM, which also employs clustering operations in the model.

First, we examine the effectiveness of federated learning training. As shown in Fig. 7, we conducted tests on both the VSM and APR datasets. The results show that the error fluctuation in the loss curve is more pronounced in the VSM dataset, while it remains relatively stable in the APR dataset. After 1000 iterations, the loss of PFL-MCL on the VSM dataset drops below 0.1, while other methods still fluctuate around 0.2. For the APR dataset, all methods demonstrate more consistent performance, with the loss values converging to within 0.1 after 1000 iterations.

Furthermore, we employ PCA visualization to analyze the clustering results, conducting tests on both the VSM and APR datasets. As show in Fig. 8, the clusters between classes illustrated by PCA are more distinct in the APR dataset with higher boundary differentiation. We can conclude that the clustering performance of both methods on the APR dataset surpasses that on the VSM dataset. On the other hand, the clustering results on the VSM dataset exhibit more overlap between categories since online task is more challenging. Despite this, when examining the results from PCA, it is evident that the clustering operations within PFL-MCL create more distinct boundaries between different classes, resulting in a clearer separation and more pronounced distances between categories. These results demonstrate the effectiveness of the proposed clustering mechanism in the federated learning process.

D. Evaluation on Recommendation Effectiveness

The varying lengths of candidate item lists may impact on the recommendation result significantly. Generally, a higher of candidate item implies a larger pool of potential items to choose from, which can make the recommendation task more challenging due to the increased number of choices.. Table II presents the results of all methods across different lengths of candidate items, M ranging from 5 to 30. The offline evaluation metrics Precision, Recall, and F1-Measure are applied for the comparison.

As shown in the table, the results from the charts show that the increasing number of candidate items M leads to an observable rise in Precision for all the methods. However, it is worth noting that Recall sees a slight variation as M increases from 5 to 15, but significantly drops when M keeps increasing to 30. This leads to an overall decrease in F1-Measure performance at M=30. First, we observe that the local model NonFed performs worse than the others for all the cases. It may be because the well-designed and finely-tuned local model with a local dataset is hardly suitable for complex real-world recommendation tasks. Second, compared with all the baseline models in all three cases, the federated learning model with clustering scheme FeSEM and PFL-MCL achieves better performance in all the metrics. Third, compared with the federated learning model FeSEM, our model achieves most 8.1% Precision, 1.8% Recall, and 5.0% F-Measure lift at M=15, and 4.8% Precision, 24.2% Recall, and 16.0% F1-Measure lift at M=30. Fourth, considering the ability to handle multi-modality when compared with MBT, the proposed federated learning scheme PFL-MCL brings 31.7%, 16.7% and 21.1% absolute gain on F1-Measure for three cases respectively.

Furthermore, we also evaluate the online performance on VSM platform with M varing from 5 to 30 as well. Table III shows the comparison results for all methods under different M. It shows that the proposed PFL-MCL outperforms other baselines in all cases in terms of the CTR metric the CTR metric. The results of online metric CTR in Table III for achieving 51% on $M = 5$ and 21% on $M = 30$ for our method indicate that PFL-MCL can support more precise items prediction in VSM platform.

VI. CONCLUSION

In this paper, to enrich the realistic and immersed experience in Metaverse-enabled smart applications, we proposed the so-called PFL-MCL model, which could better analyze users' multi-platform or cross-space data from multi-modalities for more efficient communication and networking in human-centric Metaverse over 5G and beyond networks.

A PFL framework was newly designed considering the unique characteristics of Metaverse data, which included a multi-center aggregation structure with a two-stage iterative clustering scheme in global, and a hierarchical neural network structure with a MCL scheme in local. The local training model could be further divided into a personalized module and a federated module. Specifically, a personalized multi-modal fusion network was constructed, in which a hierarchical shift-window attention mechanism was developed to effectively reduce the feature dimensions when fusing users' multi-modal input data, while a so-called bridge attention mechanism was devised to refine the cross-modal fusion from heterogeneous data with less computational cost. Moreover, a MCL scheme was improved with an embedding layer in local to reduce the communication cost and speed up the model convergence,

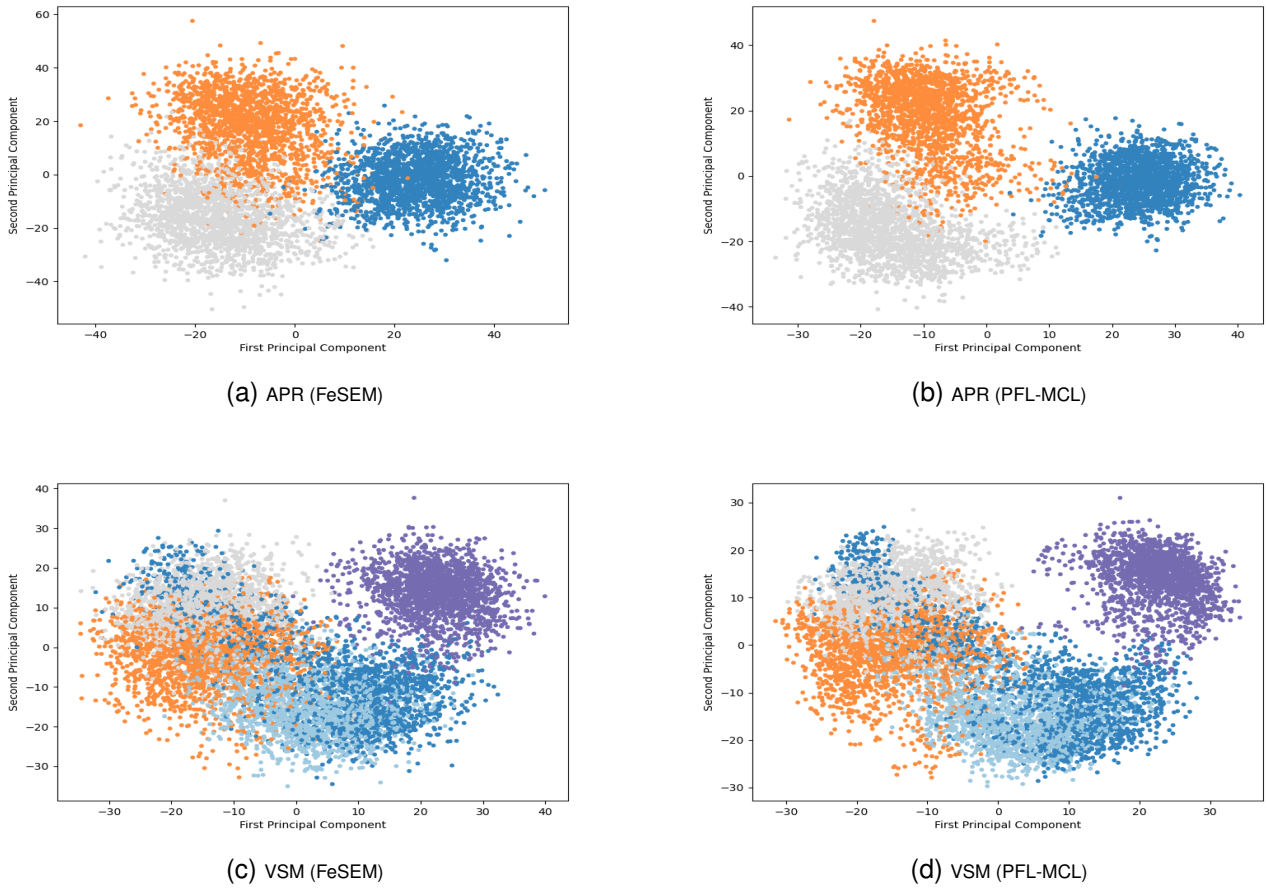

Fig. 8. The PCA visualzation of the clustering result using (a) FeSEM and (b) PFL-MCL on the APR dataset, and using (c) FeSEM and (d) PFL-MCL on the VSM dataset.

TABLE II
OFFLINE COMPARISONS ON DIFFERENT METHODS WITH DIFFERENT M

Method	$M = 5$			$M = 15$			$M = 30$		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
NonFed	52.3%	81.8%	63.8%	56.4%	85.6%	68.0%	78.0%	48.6%	59.9%
FedAvg	51.7%	81.4%	63.2%	58.1%	86.1%	69.4%	72.1%	50.9%	59.8%
MBT	53.3%	82.6%	64.8%	65.2%	86.8%	74.5%	78.4%	51.7%	62.3%
Moon	54.8%	84.4%	66.5%	62.1%	86.3%	72.2%	79.5%	52.6%	63.3%
FeSEM	66.2%	87.3%	75.3%	78.2%	87.9%	82.8%	84.6%	52.9%	65.1%
PFL-MCL	**79.9%**	**91.5%**	**85.3%**	**84.5%**	**89.5%**	**86.9%**	**88.7%**	**65.7%**	**75.5%**

TABLE III
ONLINE COMPARISONS ON DIFFERENT METHODS WITH DIFFERENT M

Method	CTR		
	M=5	M=15	M=30
NonFed	37%	18%	9%
MBT	39%	26%	13%
FedAvg	39%	28%	13%
Moon	41%	27%	15%
FeSEM	46%	35%	18%
PFL-MCL	**51%**	**37%**	**21%**

which could also make the delivered global model better adapt to the local personality. A two-stage iterative clustering algorithm was designed in global to realize a more precise initialization with dynamically updated multiple centers for personalized global aggregation. Experiments were conducted using two different real-world datasets, and evaluations compared with five baseline methods demonstrated the outstanding results of our proposed model in more efficient learning performance and recommendation accuracy, which could be applied in human-centric Metaverse environments with a fine-grain personalized training strategy.

In future studies, we will go further to conduct more evaluations in more complex situations to improve our model and algorithm with better accuracy and efficiency for more Metaverse-enabled smart systems and applications.

VII. Acknowledgment

This work was supported in part by the Grants-in-Aid for Scientific Research (C) from Japan Society for the Promotion

of Science (JSPS) under Grant 23K11064, in part by the National Natural Science Foundation of China under Grant 62072171 and Grant 72091515, in part by 2022 and 2023 Waseda University Grants for Special Research Projects under Grant 2022R-036 and Grant 2023C-216, and in part by 2023-2025 Shenzhen Science and Technology Program under Grant GJHZ20220913144201002.

REFERENCES

[1] Fengxiao Tang, Xuehan Chen, Ming Zhao, and Nei Kato. The roadmap of communication and networking in 6g for the metaverse. *IEEE Wireless Communications*, 2022.

[2] Bimal Ghimire and Danda B Rawat. Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things. *IEEE Internet of Things Journal*, 9(11):8229–8249, 2022.

[3] Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data. *arXiv preprint arXiv:1811.11479*, 2018.

[4] Qiong Wu, Xu Chen, Zhi Zhou, and Junshan Zhang. Fedhome: Cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Transactions on Mobile Computing*, 21(8):2818–2832, 2020.

[5] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[6] Hai Jin, Dongshan Bai, Dezhong Yao, Yutong Dai, Lin Gu, Chen Yu, and Lichao Sun. Personalized edge intelligence via federated self-knowledge distillation. *IEEE Transactions on Parallel and Distributed Systems*, 34(2):567–580, 2022.

[7] Xiaokang Zhou, Wei Liang, Jianhua Ma, Zheng Yan, I Kevin, and Kai Wang. 2d federated learning for personalized human activity recognition in cyber-physical-social systems. *IEEE Transactions on Network Science and Engineering*, 9(6):3934–3944, 2022.

[8] Arvin Tashakori, Wenwen Zhang, Z Jane Wang, and Peyman Servati. Semipfl: personalized semi-supervised federated learning framework for edge intelligence. *IEEE Internet of Things Journal*, 2023.

[9] Yixuan Li, Xiaoqi Qin, Hao Chen, Kaifeng Han, and Ping Zhang. Energy-aware edge association for cluster-based personalized federated learning. *IEEE Transactions on Vehicular Technology*, 71(6):6756–6761, 2022.

[10] Xianting Huang, Jing Liu, Yingxu Lai, Beifeng Mao, and Hongshuo Lyu. Eefed: Personalized federated learning of execution&evaluation dual network for cps intrusion detection. *IEEE Transactions on Information Forensics and Security*, 18:41–56, 2022.

[11] Farzan Farnia, Amirhossein Reisizadeh, Ramtin Pedarsani, and Ali Jadbabaie. An optimal transport approach to personalized federated learning. *IEEE Journal on Selected Areas in Information Theory*, 3(2):162–171, 2022.

[12] Xingjian Cao, Gang Sun, Hongfang Yu, and Mohsen Guizani. Perfedgan: Personalized federated learning via generative adversarial networks. *IEEE Internet of Things Journal*, 10(5):3749–3762, 2022.

[13] Jed Mills, Jia Hu, and Geyong Min. Multi-task federated learning for personalised deep neural networks in edge computing. *IEEE Transactions on Parallel and Distributed Systems*, 33(3):630–641, 2021.

[14] Hongzheng Yu, Zekai Chen, Xiao Zhang, Xu Chen, Fuzhen Zhuang, Hui Xiong, and Xiuzhen Cheng. Fedhar: Semi-supervised online learning for personalized federated human activity recognition. *IEEE Transactions on Mobile Computing*, 2021.

[15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[16] Man Wu, Shirui Pan, and Xingquan Zhu. Attraction and repulsion: Unsupervised domain adaptive graph contrastive learning network. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(5):1079–1091, 2022.

[17] Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5549–5560, 2022.

[18] Noureddine Kermiche. Contrastive hebbian feedforward learning for neural networks. *IEEE transactions on neural networks and learning systems*, 31(6):2118–2128, 2019.

[19] Jinqiang Wang, Tao Zhu, Liming Chen, Huansheng Ning, and Yaping Wan. Negative selection by clustering for contrastive learning in human activity recognition. *IEEE Internet of Things Journal*, 2023.

[20] Zhonghang Zhu, Lequan Yu, Wei Wu, Rongshan Yu, Defu Zhang, and Liansheng Wang. Murcl: Multi-instance reinforcement contrastive learning for whole slide image classification. *IEEE Transactions on Medical Imaging*, 2022.

[21] Yixin Liu, Zhao Li, Shirui Pan, Chen Gong, Chuan Zhou, and George Karypis. Anomaly detection on attributed networks via contrastive self-supervised learning. *IEEE transactions on neural networks and learning systems*, 33(6):2378–2392, 2021.

[22] Jinhua Zhu, Yingce Xia, Lijun Wu, Jiajun Deng, Wengang Zhou, Tao Qin, Tie-Yan Liu, and Houqiang Li. Masked contrastive representation learning for reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3421–3433, 2022.

[23] Dongxiao He, Chundong Liang, Cuiying Huo, Zhiyong Feng, Di Jin, Liang Yang, and Weixiong Zhang. Analyzing heterogeneous networks with missing attributes by unsupervised contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[24] Minrui Xu, Wei Chong Ng, Wei Yang Bryan Lim, Jiawen Kang, Zehui Xiong, Dusit Niyato, Qiang Yang, Xuemin Sherman Shen, and Chunyan Miao. A full dive into realizing the edge-enabled metaverse: Visions, enabling technologies, and challenges. *IEEE Communications Surveys & Tutorials*, 2022.

[25] Zhen Meng, Changyang She, Guodong Zhao, and Daniele De Martini. Sampling, communication, and prediction co-design for synchronizing the real-world device and digital model in metaverse. *IEEE Journal on Selected Areas in Communications*, 41(1):288–300, 2022.

[26] Yue Han, Dusit Niyato, Cyril Leung, Dong In Kim, Kun Zhu, Shaohan Feng, Xuemin Shen, and Chunyan Miao. A dynamic hierarchical framework for iot-assisted digital twin synchronization in the metaverse. *IEEE Internet of Things Journal*, 10(1):268–284, 2022.

[27] Muhammet Deveci, Dragan Pamucar, Ilgin Gokasar, Mario Köppen, and Brij B Gupta. Personal mobility in metaverse with autonomous vehicles using q-rung orthopair fuzzy sets based opa-rafsi model. *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[28] Yuna Jiang, Jiawen Kang, Dusit Niyato, Xiaohu Ge, Zehui Xiong, Chunyan Miao, and Xuemin Shen. Reliable distributed computing for metaverse: A hierarchical game-theoretic approach. *IEEE Transactions on Vehicular Technology*, 72(1):1084–1100, 2022.

[29] Minjuan Wang, Haiyang Yu, Zerla Bell, and Xiaoyan Chu. Constructing an edu-metaverse ecosystem: A new and innovative framework. *IEEE Transactions on Learning Technologies*, 15(6):685–696, 2022.

[30] Gaurang Bansal, Karthik Rajgopal, Vinay Chamola, Zehui Xiong, and Dusit Niyato. Healthcare in metaverse: A survey on current metaverse applications in healthcare. *Ieee Access*, 10:119914–119946, 2022.

[31] Yuzheng Ren, Renchao Xie, Fei Richard Yu, Tao Huang, and Yunjie Liu. Quantum collective learning and many-to-many matching game in the metaverse for connected and autonomous vehicles. *IEEE Transactions on Vehicular Technology*, 71(11):12128–12139, 2022.

[32] Haoran Shi, Guanjun Liu, Kaiwen Zhang, Ziyuan Zhou, and Jiacun Wang. Marl sim2real transfer: Merging physical reality with digital virtuality in metaverse. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(4):2107–2117, 2022.

[33] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, volume 34, pages 14200–14213, 2021.

[34] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10708–10717, 2021.

[35] Guodong Long, Ming Xie, Tao Shen, Tianyi Zhou, Xianzhi Wang, and Jing Jiang. Multi-center federated learning: clients clustering for better personalization. *World Wide Web*, 26(1):481–500, 2023.

[36] Bickey Kumar Shah, Amar Kumar Jaiswal, Anshul Shroff, Ashutosh Kumar Dixit, Om Nath Kushwaha, and Nisha Kumari Shah. Sentiments detection for amazon product review. In *2021 International conference on computer communication and informatics (ICCCI)*, pages 1–6. IEEE, 2021.

[37] H Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2:2, 2016.