Information Fusion 126 (2026) 103578

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

# Seeking personality yet commonality via adversarial learning to enhance heterogeneous multimodal collaboration

Zhuojia Wu [a], Qi Zhang [a], Duoqian Miao [a,*], Xuerong Zhao [b], Guangyin Bao [a], Liang Hu [a], Kun Yi [c], Yu Zhou [c]

[a] The Department of Computer Science and Technology, Tongji University, Shanghai, 201804, China
[b] The College of Information Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, 201418, China
[c] North China Institute of Computing Technology, Beijing, 100083, China

ARTICLE INFO

*Keywords:*
Multimodal federated learning
Privacy preservation
Affective computing
Generative adversarial network
Heterogeneous multimodal collaboration

ABSTRACT

Online platforms amass extensive client data, providing valuable multimodal information for collaborative learning across platforms. With the increasing diversity of online content, client data undergoes significant heterogeneities, such as domain shifts, modality gaps, and task drifts, increasing gradient conflicts and negative transfer effects in collaborative learning. Consequently, there is an urgent need to design a robust method for cross-client knowledge sharing for heterogeneous multimodal collaboration. In this paper, we propose a novel distributed collaborative learning framework, called HeMuGAN, leveraging customized GANs to facilitate personalized knowledge sharing among heterogeneous clients. Unlike existing parameter-based federated aggregation models, HeMuGAN deploys generators and discriminators across participating clients, enabling each client to learn personalized knowledge from the data space of other heterogeneous clients while also serving as a knowledge provider to others. Extensive experiments show that HeMuGAN consistently achieves superior knowledge sharing for multimodal collaborative learning, significantly improving client downstream tasks. Further privacy analysis also verifies its capability in privacy preservation.

## 1. Introduction

Today, numerous online platforms provide users with convenient services while amassing vast reservoirs of client data. Utilizing these scattered data resources effectively is paramount for elevating service quality and enriching user experiences, increasingly attracting devotion to collaboration among platforms. Consequently, pursuing meaningful knowledge sharing through collaborative learning is a common goal and is increasingly focused in industry and academia [1,2].

However, direct data sharing for mutual access can lead to privacy risks. Moreover, with the increasing diversification of online services and user-generated content, platforms across different domains often contain a variety of modalities and are oriented towards specific downstream tasks [3]. For instance, e-commerce platforms analyze user text feedback to pinpoint product flaws and enhance their offerings; customer service systems gauge user sentiments through voice interactions to refine services; and social media screens short videos to weed out extreme comments. Such commonplace heterogeneous domains, modalities, and tasks among platforms (clients) present substantial obstacles to conventional collaborative learning, resulting in corresponding significant challenges of domain shifts, modality gaps, and task drifts to knowledge sharing (i.e., commonality learning). This motivates us to rethink the significance and necessity of specific knowledge (i.e., personalized learning) of each client, raising an interesting question: ***How can each client effectively acquire the personalized knowledge to enhance collaborative learning?***

Federated Learning (FL) [4–7], as a prominent Distributed Collaborative Learning (DCL) framework [8,9], enables knowledge sharing by aggregating model parameters trained on local data from different clients at a central server. Traditional-FL presupposes uniform model architectures across all clients, confining participating clients to the same data modality and unified tasks [10]. Recent Multimodal-FL introduces specific local model architectures, seeking to facilitate the sharing of domain, modality, and task-agnostic general knowledge [10,11]. Typically, it involves decomposing the model into multiple sub-blocks and consolidating parameters for these shared sub-blocks across clients [9]. However, these methods still encounter limitations regarding the structures of local models. Herein, sophisticated disentanglement strategies are necessary to pinpoint additional shared

[*] Corresponding author.
*E-mail addresses:* wuzhuojia@tongji.edu.cn (Z. Wu), zhangqi_cs@tongji.edu.cn (Q. Zhang), dqmiao@tongji.edu.cn (D. Miao), xrzhao@tongji.edu.cn (X. Zhao), baogy@tongji.edu.cn (G. Bao), lianghu@tongji.edu.cn (L. Hu), kunyi.cn@gmail.com (K. Yi), yuzh_a@hotmail.com (Y. Zhou).

https://doi.org/10.1016/j.inffus.2025.103578
Received 24 April 2025; Received in revised form 18 June 2025; Accepted 21 July 2025
Available online 30 July 2025
1566-2535/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

sub-blocks to enhance commonality learning [12]. The latest methods divide each client's local model into two distinct components: one for acquiring shared knowledge and the other for optimizing local downstream tasks [12]. This approach effectively harmonizes global common knowledge and personalized local insights, enabling clients to build flexible local models without intricate disentanglement computations.

Although recent Multimodal-FL demonstrates the potential for knowledge sharing across heterogeneous clients, it follows the parameter aggregation scheme of Traditional-FL. This, however, leads to the following limitations: ***(1) Insufficient Personalized Fulfillment***. Only a single global model is produced with primarily common knowledge. It often fails to cater to the personalized knowledge needs of heterogeneous clients; ***(2) Inefficient Knowledge Integration***. Traditional-FL frequently experiences gradient conflicts during parameter aggregation due to client heterogeneity, such as label distribution skew [13,14]. This leads to negative knowledge transmission and, in some cases, inferior performance than independent learning. Multimodal-FL further exacerbates the issue with its diverse heterogeneity. *These observations inspire us with a new approach to knowledge sharing among heterogeneous clients, with a focus on ensuring confidentiality while attaining effective and personalized knowledge sharing.*

Intuitively, if each client can directly acquire knowledge from the data of other clients, it is an optimal choice for personalized knowledge sharing. In this setting, each client essentially owns data privacy yet possesses complete knowledge, where clients can flexibly fashion models and extract the most relevant knowledge. However, the intricate heterogeneity presents significant challenges to achieving optimal personalized learning due to data inconsistency, modality gaps, and task shifts, additionally with real-world noises. Accordingly, to enable each client to flexibly and efficiently learn personalized knowledge from other clients, two pivotal issues necessitate resolution: ***(1) mitigating the risks of privacy leakage***; ***(2) eliminating the obstacles of knowledge sharing***.

In light of the above discussion, we propose a novel DCL framework, called HeMuGAN, leveraging customized GANs to facilitate personalized knowledge sharing among heterogeneous clients. Specifically, in HeMuGAN, each client simultaneously utilizes local and other participating clients' training samples, with the latter consisting of de-identified representations extracted from the model's intermediate layers rather than raw data, concentrating on learning local personalized tasks. Unlike existing parameter aggregation schemes, this sharing pattern allows each client to acquire knowledge from other participants without accessing their raw data, thereby preserving privacy. Furthermore, we deploy competing generators and discriminators on all clients that operate on the local representations and external representations from other clients. The generators and discriminators enable cross-domain transfer, cross-modality reconstruction, and cross-task adaptation, overcoming the barriers to knowledge sharing posed by client heterogeneity. As a result, the proposed HeMuGAN ensures that, while preserving privacy, each client's local model can efficiently acquire personalized knowledge from other clients for heterogeneous multimodal collaboration. Our contributions are summarized below:

- We propose a novel DCL framework, HeMuGAN, for heterogeneous multimodal collaboration, which, to our best knowledge, marks the first to enable personalized knowledge learning for each client while preserving privacy.
- We tailor GANs for collaborative learning among heterogeneous clients to achieve cross-domain transform, cross-modality reconstruction, and cross-task adaptation, paving the way for efficient knowledge sharing.
- Extensive experiments under various heterogeneous scenarios verify the effectiveness of personalized multimodal collaborative learning. Further analysis confirms the complete privacy preservation of HeMuGAN.

## 2. Related work

### 2.1. Multimodal federated learning

Traditional-FL aims to train a global model using data distributed across different clients. The underlying assumption is that the data across clients exhibit only statistical heterogeneity, meaning they are entirely consistent in terms of domain, modality, and task. In contrast, Multimodal-FL seeks to enable distributed collaborative learning among clients with heterogeneous domains, modalities, and tasks, improving the performance of personalized models on each client through knowledge sharing. Most existing Multimodal-FL methods inherit the parameter-based aggregation for knowledge sharing from Traditional-FL. Yang et al. [15] proposed decomposing local models into modality-agnostic and modality-specific sub-blocks, where the former is shared across all clients, while the latter is shared only among clients with the same modality. Based on this, FedMSplit [9] further uses a dynamic cross-client multi-view graph structure to capture inter-block correlations. Furthermore, Cross-Modal Meta Consensus (CMMC) [3] was proposed to address the knowledge asymmetry issue by maximizing positive knowledge sharing while minimizing negative transfer. In addition, Yu et al. [16] proposed a contrastive federated framework based on knowledge distillation. The method addresses model drift and task drift by introducing a global–local cross-modality contrastive strategy, facilitating knowledge sharing among clients.

### 2.2. Affective computing

Affective computing [17] aims to enable intelligent systems to recognize and understand human emotions, thereby enhancing the accuracy and richness of human–computer interactions. With advancements in technology, it has wide application in e-commerce, customer service, and social media [18–20]. The data to be analyzed and the downstream tasks have also become more diverse, including aspect-level sentiment analysis of text reviews [21], emotion recognition in spoken conversations [22], and opinion detection in video news content [23,24]. Platforms are eager to share raw data to improve the performance of local systems. However, direct sharing is not feasible due to existing sensitive information, such as personal identifiers and interaction records. These facts highlight that affective computing is a field requiring privacy preservation, while also facing widespread heterogeneity across domains, modalities, and tasks among different clients. Moreover, different affective computing tasks often involve shared generalizable knowledge [25], which enables to assess the impact of knowledge sharing through the performance improvements observed across tasks. Consequently, we develop a simulation environment based on affective computing to investigate more efficient DCL frameworks.

### 2.3. Generative adversarial networks

Generative Adversarial Networks (GANs) [26] were initially introduced for generating handwritten digit images. Since then, numerous improvements have focused on enhancing the quality of generated images [27–29]. Specifically, Conditional GANs (cGANs) [27] introduced conditional supervision into adversarial learning, allowing the generator to produce samples related to a given condition. Additionally, some studies have explored the application of GANs to solve other tasks, such as domain adaptation for transferring knowledge between source and target domains [30,31]. GANs have also been used in areas like data augmentation [32], semi-supervised learning [33], text generation [34], and others. In our work, customized GANs are used to overcome the barriers to knowledge sharing caused by heterogeneities among clients, representing efforts to apply GANs in distributed collaborative learning.

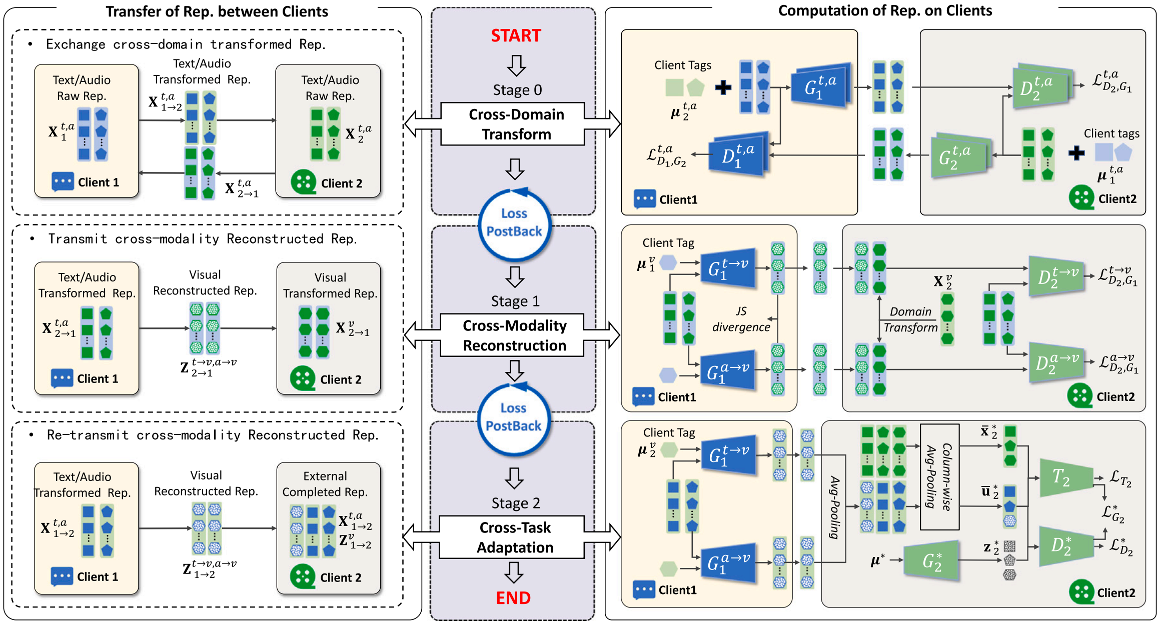


**Fig. 1.** Overall distributed workflow of our proposed HeMuGAN. Assume that the original data of Client1 consists of text and audio, while Client2 includes text, audio, and visual (*Modality Gap*). Moreover, even within the same modality across clients, there is a distribution discrepancy (*Domain-Shift*) and each client has personalized local tasks (*Task Drift*). Specifically, different shapes denote different modality representations (Rep.), while varying background colors indicate different domains. The color changes represent domain transformations.

## 3. Methodology

### 3.1. Problem statement

Given $N$ independent clients, each client $i \in \{1, 2, \dots, N\}$ has its local dataset $C_i$, which is characterized by the domain distribution $\mathcal{D}_i$, modality set $\mathcal{M}_i$, and task space $\mathcal{T}_i$. In particular, we consider the three most common data modalities: text ($t$), audio ($a$), and visual ($v$), such that $\mathcal{M}_i \subseteq \{t, a, v\}$. Given another client $j \in \{1, 2, \dots, N\}$ and $j \neq i$, the multiple non-statistical heterogeneities between clients $i$ and $j$ are defined as : (1) *Domain Shift*, i.e., $\mathcal{D}_i \nsim \mathcal{D}_j$. In this context, it refers to the distribution differences in the same modality data arising from domain variations. (2) *Modality Gap*, i.e., $\mathcal{M}_i \nsubseteq \mathcal{M}_j$ and $\mathcal{M}_j \nsubseteq \mathcal{M}_i$, indicating that the input spaces of the local models for clients $i$ and $j$ differ. (3) *Task Drift*, i.e., $\mathcal{T}_i \ncong \mathcal{T}_j$, signifying that the output spaces of the models for clients $i$ and $j$ are different. The heterogeneity requires clients to develop personalized models, leading to inefficiencies in knowledge sharing through parameter aggregation.

In HeMuGAN, knowledge sharing among clients is achieved by transferring the de-identified representations from the intermediate layer outputs of the local models. Furthermore, upon receiving external representations from other clients, each client can construct personalized fusion layers and decoders based on its local tasks, thereby learning personalized knowledge while avoiding direct access to the original data. For each sample in dataset $C_i$, $\mathbf{X}_i^{\mathcal{M}_i} = [\mathbf{x}_i^{\mathcal{M}_i}]_{0:L}$ represents the raw embedding representation sequence of its different modalities, which corresponds to the output of the shallow encoder. Here, $L$ is the sequence length. In addition, all components in HeMuGAN use the vanilla Transformer [35] as the backbone, except for the prediction layer.

### 3.2. Overall workflow

As depicted in Fig. 1, the workflow of HeMuGAN involves distributed collaborative learning between two heterogeneous multimodal clients. The representation transfer between clients is structured into three distinct stages:

(1) At Stage 0, clients exchange cross-domain transformed representations. This exchange relies on the existence of common modalities between the two clients. For instance, both Client1 and Client2 possess text and audio modalities, and they exchange their respective text and audio representations after the cross-domain transformation. Each client then computes the discriminator loss using its local raw representations and the received external representations, and returns the generator loss to the sender. This process guides the sender's generator to produce external representations that are consistent with the receiver's local representation distribution.

(2) At Stage 1, clients transmit cross-modality reconstructed representations. This transmission depends on the other client possessing the locally missing modality. For example, Client2 possesses the visual modality, while Client1 does not. Therefore, Client1 generates the cross-modality reconstructed visual representation based on the external representation received from Client2 in the previous stage and sends it back to Client2. Next, Client2 calculates the discriminator loss based on the real visual representation (with cross-domain transformation performed first) and the received reconstructed representation. It then returns the generator loss to the sender to guide the generator in producing a more realistic reconstructed representation.

(3) At Stage 2, clients re-transmit cross-modality reconstructed representations. This transmission also relies on the other client possessing the locally missing modality. Using the same example, Client1 again generates reconstructed visual representations and sends them to Client2. Notably, unlike Stage 1, the generator is now frozen and no adversarial loss is computed. Instead, the reconstructed representations are integrated into Client2's external representations to fill in the missing modalities. This results in both modality- and domain-aligned external representations, from which Client2 can extract personalized knowledge via its local task-adaptive GANs.

## 4. GANs for efficient knowledge sharing

### 4.1. Cross-domain transform

Due to domain shift, we introduce adversarial learning between client $i$ and client $j$ for cross-domain transform of external representations, ensuring that the external representations received by client $i$

**Algorithm 1:** Multi-Stage Layerwise Training of HeMuGAN

```
Input: Multiple training clients $i, j \in \{1, 2, ..., N\}$, where $i \neq j$; modality sets $\mathcal{M}_i$ and $\mathcal{M}_j$; training iterations $I_0, I_1, I_2$.
foreach client $i, j$ in parallel do
    if $m \in \mathcal{M}_i \cap \mathcal{M}_j$ and $m \neq \emptyset$ then
        // Stage 0: Cross-Domain Transform
        Initialize $\mathbf{W}^m_{G_i}$, $\mathbf{W}^m_{D_i}$, $\mathbf{W}^m_{G_j}$, and $\mathbf{W}^m_{D_j}$;
        for $0 \to I_0$ do
            $\triangleright$ $\mathbf{X}^m_{j\to i}, \mathbf{X}^m_{i\to j}$ are computed as Eq.1, then Client $i$ sends $\mathbf{X}^m_{i\to j}$ to Client $j$ and receives $\mathbf{X}^m_{j\to i}$;
            $\mathcal{L}^m_{D_i}$ and $\mathcal{L}^m_{D_j}$ are computed as Eq.2, with $\mathbf{W}^m_{D_i}$ and $\mathbf{W}^m_{D_j}$ updated accordingly;
            $\triangleright$ $\mathcal{L}^m_{G_i}$ and $\mathcal{L}^m_{G_j}$ are computed as Eq.3, then Client $i$ sends $\mathcal{L}^m_{G_j}$ to Client $j$ and receives $\mathcal{L}^m_{G_i}$;
            $\mathbf{W}^m_{G_i}$ and $\mathbf{W}^m_{G_j}$ are updated accordingly;
        // Stage 1: Cross-Modality Reconstruction
        if $n \in \mathcal{M}_j \setminus \mathcal{M}_i$ and $n \neq \emptyset$ then
            Initialize $\mathbf{W}^{m\to n}_{G_i}$, $\mathbf{W}^{m\to n}_{D_j}$, and freeze the parameters from Stage 0 ;
            for $0 \to I_1$ do
                $\triangleright$ $\mathbf{Z}^{m\to n}_{j\to i}$ is computed as Eq.4, then Client $i$ sends $\mathbf{Z}^{m\to n}_{i\to j}$ to Client $j$;
                $\mathcal{L}^{m\to n}_{D_j}$ is computed as Eq.5, with $\mathbf{W}^{m\to n}_{D_j}$ updated;
                $\triangleright$ $\mathcal{L}^{m\to n}_{G_i}$ is computed as Eq.6, then Client $j$ sends $\mathcal{L}^{m\to n}_{G_i}$ to Client $i$ with $\mathbf{W}^{m\to n}_{G_i}$ updated ;
            $\triangleright$ $\mathbf{Z}^{m\to n}_{i\to j}$ is calculated as Eq.4, then Client $i$ sends $\mathbf{Z}^{m\to n}_{i\to j}$ to Client $j$;
        // Stage 2: Cross-Task Adaptation
        Initialize $\mathbf{W}^*_{D_k}$, $\mathbf{W}_{T_k}$, $\mathbf{W}^*_{G_k}$, where $k \in \{i, j\}$, and freeze the parameters from Stage 0 and 1 ;
        for $0 \to I_2$ do
            $\mathcal{L}^*_{D_k}, \mathcal{L}_{T_k}, \mathcal{L}^*_{G_k}$ are computed as Eq.8-10, with $\mathbf{W}^*_{D_k}$, $\mathbf{W}_{T_k}$, $\mathbf{W}^*_{G_k}$ updated accordingly;
```

from client $j$ are consistent with its local distribution, enabling efficient knowledge learning by the local model. Specifically, the GAN designed for cross-domain transform consists of a generator and a discriminator. The generator is placed on the sharing side, i.e., client $j$, while the discriminator is positioned on the receiving side, i.e., client $i$. The transform process of the representation $\mathbf{X}^m_j$ by the generator is formalized as:

$$\mathbf{X}^m_{j\to i} = G^m_j(\mathbf{X}^m_j, \boldsymbol{\mu}^m_i; \mathbf{W}^m_{G_j}),\ \ m \in \mathcal{M}_i \cap \mathcal{M}_j \tag{1}$$

where $\mathcal{M}_i$ and $\mathcal{M}_j$ represent the sets of data modalities possessed by client $i$ and $j$, respectively. $\mathbf{W}^m_{G_j}$ is the set of trainable parameters. $\boldsymbol{\mu}^m_i$ is a client tag vector, serving as conditional information, and is concatenated at the beginning of the input sequence $\mathbf{X}^m_j$ to guide the generator in transforming $\mathbf{X}^m_j$ to the target domain.

Then, the transformed representation $\mathbf{X}^m_{j\to i}$ is sent to the target client $i$, in which it is combined with the client's own representation $\mathbf{X}^m_i$ and fed into the discriminator $D^m_i$. The discriminator updates its parameters based on the following binary cross-entropy loss:

$$\begin{aligned}\mathcal{L}^m_{D_i} &= -\,log\,D^m_i(\mathbf{X}^m_i, \boldsymbol{\mu}^m_i; \mathbf{W}^m_{Di}) \\ &\quad -\,log(1 - D^m_i(\mathbf{X}^m_{j\to i}, \boldsymbol{\mu}^m_i; \mathbf{W}^m_{Di}))\end{aligned} \tag{2}$$

where $\mathbf{W}^m_{Di}$ is the set of trainable parameters. Meanwhile, the discriminator also provides supervision information to the generator. The generator's loss function is as follows:

$$\mathcal{L}^m_{G_j} = -log\,D^m_i(\mathbf{X}^m_{j\to i}, \boldsymbol{\mu}^m_i; \mathbf{W}^m_{Di}) \tag{3}$$

where $\mathcal{L}^m_{G_j}$ is returned to Client $j$, guiding the generator $G^m_j$ to optimize its parameters such that the external representations it produces are difficult for the discriminator to distinguish between domains.

### 4.2. Cross-modality reconstruction

For the modality gap among clients, we treat it as an issue of missing modalities in some clients and deploy customized distributed GANs to reconstruct the missing modality for these clients. Specifically, suppose that client $i$ has missing modalities compared to client $j$. In this case, Distributed GANs are employed between the clients to reconstruct the missing modalities for client $i$. The generators on client $i$ use the received external representation $\mathbf{X}^m_{j\to i}$ as conditional information to generate cross-modality reconstructed representations that align semantically, serving as complemented features for the missing modalities [27].

Additionally, to maintain the contextual relevance of elements within the generated representation sequence, the generators produce feature representation at each timestamp in an autoregressive manner [6]. It is formalized below:

$$\begin{aligned}[\mathbf{z}^{m\to n}_{j\to i}]_t &= G^{m\to n}_i([\mathbf{z}^{m\to n}_{j\to i}]_{0:t-1}, \mathbf{X}^m_{j\to i}; \mathbf{W}^{m\to n}_{G_i}), \\ &\quad m \in \mathcal{M}_i \cap \mathcal{M}_j,\ n \in \mathcal{M}_j \setminus \mathcal{M}_i\end{aligned} \tag{4}$$

with trainable parameters $\mathbf{W}^{m\to n}_{G_i}$ and $[\mathbf{z}^{m\to n}_{j\to i}]_0 = \boldsymbol{\mu}^n_i$. The representation $\mathbf{Z}^{m\to n}_{j\to i}$ is sent to Client $j$, with the real representation of the missing modalities fed into the discriminator for prediction. The discriminator's loss is as follows:

$$\begin{aligned}\mathcal{L}^{m\to n}_{D_j} &= -(\sum_{t=0}^{L} log\,D^{m\to n}_j([\mathbf{x}^n_{j\to i}]_t, \mathbf{X}^m_{j\to i}; \mathbf{W}^{m\to n}_{D_j}) \\ &\quad -log(1 - D^{m\to n}_j([\mathbf{z}^{m\to n}_{j\to i}]_t, \mathbf{X}^m_{j\to i}; \mathbf{W}^{m\to n}_{D_j})))\end{aligned} \tag{5}$$

where $\mathbf{W}^{m\to n}_{D_j}$ is the set of trainable parameters and $L$ is the length of reconstructed representation $\mathbf{Z}^{m\to n}_{j\to i}$.

In particular, $\mathbf{X}^n_{j\to i}$ is obtained from $\mathbf{X}^n_j$ through a non-parametric domain transform, and the domain distance between them is defined as the cosine similarity between $\mathbf{X}^m_{j\to i}$ and $\mathbf{X}^m_i$. The generator's loss is computed as follows:

$$\mathcal{L}^{m\to n}_{G_i} = -\sum_{t=0}^{L} log\,D^{m\to n}_j([\mathbf{z}^{m\to n}_{j\to i}]_t, \mathbf{X}^m_{j\to i}; \mathbf{W}^{m\to n}_{D_j}) \tag{6}$$

where $\mathcal{L}^{m\to n}_{G_i}$ is returned to client $i$ to optimize $G^{m\to n}_i$.

When one modality contains multiple reconstructed representations from other modalities, we apply Jensen–Shannon Divergence (JSD) among them to ensure their similarity [36]. The underlying principle of the GAN in Cross-Modality Reconstruction is to leverage the natural relationships between different modalities in multimodal data through adversarial learning. For instance, in multimodal sentiment analysis, even if one modality is missing, people can intuitively infer the sentiment of the missing modality based on the available modalities [6].

### 4.3. Cross-task adaptation

For client $i$, $\{\mathbf{X}_i^m, \mathbf{Z}_i^{m\to n}\}$ including the raw and reconstructed modality representations enables the utilization of complementary knowledge from different modalities. The large-scale set of external multimodal representations $\{\mathbf{X}_{j\to i}^m, \mathbf{Z}_{j\to i}^{m\to n}\}$ provides greater diversity and richer contextual information. To efficiently acquire the required personalized knowledge from external representations, we treat external representations as features of unlabeled data, generate pseudo-labels for them, and use high-confidence pseudo-labels as supervisory signals to optimize the local model.

To improve the quality of the pseudo-labels, adversarial learning is established between the generator, discriminator, and predictor on Client $i$. The generator is defined:

$$\mathbf{z}_i^* = G_i^*(\boldsymbol{\mu}^*; \mathbf{W}_{G_i}^*) \tag{7}$$

with trainable parameters $\mathbf{W}_{G_i}^*$ and $\boldsymbol{\mu}^* \sim \mathcal{N}(0,1)$.

The discriminator is to distinguish whether the input is produced by the generator. Its loss is defined below:

$$\begin{aligned}\mathcal{L}_{D_i}^* = &- log\, D_i^*(\bar{\mathbf{x}}_i^*, W_{D_i}^*) - log\, D_i^*(\bar{\mathbf{u}}_i^*, W_{D_i}^*)\\ &- log(1 - D_i^*(\mathbf{z}_i^*, W_{D_i}^*))\end{aligned} \tag{8}$$

where $\bar{\mathbf{x}}_i^*$ denotes the representation of locally labeled samples, obtained by concatenating $\{\mathbf{X}_i^m, \mathbf{Z}_i^{m\to n}\}$ with column-wise average pooling. Similarly, $\bar{\mathbf{u}}_i^*$ denotes the representation of the unlabeled sample, which is computed based on $\{\mathbf{X}_{j\to i}^m, \mathbf{Z}_{j\to i}^{m\to n}\}$.

The task predictor provides predictions for all input samples. Specifically, assuming there is a classification task, for local labeled samples, the predictor aims to minimize the loss between its predicted values and the true labels. For unlabeled samples, the predictor's objective is to enhance its confidence in the current prediction. Conversely, for samples produced by the generator, the predictor aims to reduce its confidence in the current predicted class. The overall loss for the predictor is as follows:

$$\begin{aligned}\mathcal{L}_{T_i} = &- y_i^{[\bar{\mathbf{x}}]} log T_i(\bar{\mathbf{x}}_i^*; \mathbf{W}_{T_i}) - \hat{y}_i^{[\bar{\mathbf{u}}]} log T_i(\bar{\mathbf{u}}_i^*; \mathbf{W}_{T_i})\\ &- \hat{y}_i^{[\mathbf{z}]} log(1 - T_i(\mathbf{z}_i^*; \mathbf{W}_{T_i}))\end{aligned} \tag{9}$$

where $y_i^{[\bar{\mathbf{x}}]}$ is the true labeled class of $\bar{\mathbf{x}}_i^*$, $\hat{y}_i^{[\bar{\mathbf{u}}]}$ and $\hat{y}_i^{[\mathbf{z}]}$ are the predicted classes for $\bar{\mathbf{u}}_i^*$ and $\mathbf{z}_i^*$, respectively.

The objective of the generator is to deceive the discriminator and encourage the predictor to make more confident predictions on the generated samples. Thus, its loss is defined as follows:

$$\mathcal{L}_{G_i}^* = -log\, D_i^*(\mathbf{z}^*, \mathbf{W}_{D_i}^*) - \hat{y}_i^{[\mathbf{z}]} log T_i(\mathbf{z}^*; \mathbf{W}_{T_i}) \tag{10}$$

where $\mathcal{L}_{G_i}^*$, $\mathcal{L}_{T_i}$, and $\mathcal{L}_{D_i}^*$ are alternately optimized.

The inclusion of the generator and discriminator prevents the predictor from being overconfident in making incorrect decisions on samples near class boundaries. Such decisions could be mistakenly used as reliably predicted labels, misleading the predictor during semi-supervised learning [37].

### 4.4. Multi-stage layerwise training

Multiple losses are typically combined into a unified objective through a weighted sum and optimized jointly via backpropagation. However, jointly optimizing multiple groups of adversarial losses is highly challenging. Inspired by accelerated training strategies commonly adopted in large-scale language models [38], HeMuGAN employs a multi-stage layerwise training scheme, as illustrated in Algorithm 1. Specifically, at the initial stage (`Stage 0`), HeMuGAN trains only the GAN responsible for cross-domain transform. In subsequent stages, specialized GAN modules for cross-modality reconstruction and cross-task adaptation are progressively introduced. Notably, at each stage, only the newly introduced parameters are optimized, while those from previous stages remain frozen. This incremental training strategy enables each module to specialize in its intended function, mitigating gradient conflicts and accelerating training by reducing the depth of backpropagation, given that backward propagation generally incurs higher computational costs than forward propagation.

**Table 1**
Summary statistics of four experimental datasets.

| Datasets | Train | Valid | Test | Total |
|---|---|---|---|---|
| REST14 | 3,608 | – | 1120 | 4,728 |
| MELD | 9,989 | 1109 | 2610 | 13,708 |
| MOSI | 1,284 | 229 | 686 | 2,199 |
| MOSEI | 16,326 | 1871 | 4659 | 22,856 |

## 5. Experiments

### 5.1. Datasets

We selected four datasets from different domains, modalities, and tasks to simulate a heterogeneous environment in multimodal collaboration learning. Specifically:

- **REST14** [39] is a **uni-modality** dataset (text only) composed of restaurant reviews, established for Aspect-based Sentiment Analysis (SA). Each textual review involves one or more aspects (such as food, service, and environment), with each aspect assigned a sentiment polarity label, which can be negative, neutral, or positive.
- **MELD** [20] is a **bi-modality** dataset (text and audio), consisting of multiple conversations for Emotion Recognition (ER). Each speech conversation is composed of several consecutive utterances, where each utterance is annotated with one of seven emotion categories.
- **MOSI** [18] is a **tri-modality** dataset (text, audio, and visual) composed of video clips from YouTube monologues, used for speaker Sentiment Intensity Regression (SIR). Each video clip expresses speakers' sentiment intensity towards a specific topic, represented by a continuous value in the range of [−3, +3], where −3 indicates a strong negative, and +3 indicates a strong positive.
- **MOSEI** [19] is an extended version of MOSI, containing a larger number of video clips, covering a wider range of topics, more reviews, and additional speakers, while retaining the same annotation information as MOSI.

The dataset statistics are provided in Table 1. To ensure a fair comparison, we followed the same training, validation, and test splits as those used in the baseline models [21–23].

### 5.2. Distributed setting and evaluation criteria

To validate the performance of HeMuGAN in heterogeneous scenarios, we simulate real-world distributed settings using three datasets. Each dataset is treated as a client with privacy preservation requirements, meaning that data cannot leave the local environment throughout the entire workflow. The tasks executed by the clients include three types: SA, ER, and SIR. A total of eight evaluation metrics are used, including binary classification accuracy (**Acc-2**) and **F1-Score**,

Table 2
Hyperparameter settings of HeMuGAN for REST14, MELD, and MOSEI datasets.

| Terms | REST14 | MELD | MOSEI |
|---|---|---|---|
| **Num of transformer layers/heads** | | | |
| *Encoder* | | | |
| $t$ (Text) | BERT | BERT | BERT |
| $a$ (Audio) | – | 3/1 | 3/1 |
| $v$ (Visual) | – | – | 3/2 |
| *GAN for Cross-Domain transform* | | | |
| – $G_i^t/D_i^t$ | 2/2 | 2/2 | 2/2 |
| – $G_i^a/D_i^a$ | – | 2/1 | 2/1 |
| *GAN for Cross-Modality reconstruction* | | | |
| – $G_i^{t\to a}$ | 3/1 | – | – |
| – $D_i^{t\to a}$ | – | 3/1 | 3/1 |
| – $G_i^{t\to v}$ | 3/2 | 3/2 | – |
| – $D_i^{t\to v}$ | – | – | 3/2 |
| – $G_i^{a\to v}$ | – | 3/1 | – |
| – $D_i^{a\to v}$ | – | – | 3/1 |
| **Num of perceptron layers/dimensions** | | | |
| *GAN for Cross-Task adaptation* | | | |
| – $G_i^*$ | 3/(50, 500, 1320) | 3/(50, 500, 1320) | 3/(50, 500, 1320) |
| – $D_i^*$ | 3/(1320, 500, 1) | 3/(1320, 500, 1) | 3/(1320, 500, 1) |
| – $T_i$ | 3/(1320, 500, 3) | 3/(1320, 500, 7) | 3/(1320, 500, 1) |
| **Learning rates** | | | |
| $G$ | 2e−4 | 2e−4 | 2e−4 |
| $D$ | 1e−4 | 1e−4 | 1e−4 |
| $T$ | 1e−4 | 1e−4 | 1e−5 |

three-class classification accuracy (**Acc-3**) and **Macro-F1**, seven-class classification accuracy (**Acc-7**) and weighted F1-score (**W-F1**), along with Mean Absolute Error (**MAE**) and Pearson Correlation Coefficient (**Corr**) for regression tasks. All reported experimental results are obtained by setting random seeds and repeating the training process five times, with the average results presented.

### 5.3. Hyperparameters

HeMuGAN is developed using Python 3.8.18 and PyTorch 2.0.0. Each dataset is assigned to an independent client, with the models deployed on separate RTX 4090 (24 GB) GPUs for each client. Following the previous works [23,40,41], we use the librosa library [42] to extract frame-level acoustic features, including 40-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) from the audio modality. For the visual modality, we employ a dual network combining the Multi-task Cascaded Convolutional Network (MTCNN) [43] and ResNet-101 [44], both pre-trained on the AffectNet [45] dataset, to extract 512-dimensional features related to the speaker's facial expressions and scene context for each frame of the image. Subsequently, the low-level features from each modality are fed into a modality-specific encoder, constructed with multiple transformer layers, to capture the corresponding informative raw representations. Specifically, for the text modality, we use the most representative large language model, BERT [46], to obtain the text raw representations for each sample, where each word is encoded as a 768-dimensional embedding vector. The parameters of BERT are fine-tuned together with the downstream task's loss, without requiring an additional Transformer-based text encoder. In HeMuGAN, all losses are optimized using the Adam optimizer [47], with betas set to (0.5, 0.999), and the batch size for all datasets is uniformly set to 32. The hyperparameters of the networks built on different datasets in HeMuGAN are shown in Table 2.

### 5.4. Baselines

We compare HeMuGAN with three kinds of baselines: fully independent training on each client (**Local**); addressing only statistical heterogeneity (**Traditional-FL**); and focusing on resolving the multiple non-statistical heterogeneities to promote cross-client knowledge sharing (**Multimodal-FL**). The details of the baselines are as follows:

- **SOTA Models** [21–23] include all State-Of-The-Art (SOTA) models for each downstream task, which are built on the Transformer backbone and then further improved to adapt to downstream tasks.
- **Transformer** [35] is used to encode representation sequences from different modalities, followed by column-wise average pooling of the output representations. Next, the pooled representations from different modalities are concatenated for prediction.
- **Pre-Finetuning** refers to the process where the model is first fine-tuned on datasets from other clients before being independently trained on each client's own data.
- **Fedavg** [48] is the foundational work of FL, where the global model is generated by averaging the model parameters trained locally on each client.
- **Fedproto** [49] addresses statistical heterogeneity across clients by using prototypical networks to learn and share modality prototypes across clients, guiding local model training and improving global model consistency.
- **FedGKD** [50] alleviates the issue of statistical heterogeneity across clients by utilizing knowledge distillation to regularize local model updates, ensuring more consistent global learning across diverse data distributions.
- **Meta-HAR** [51] designs a shared embedding network and is trained through a model-agnostic meta-learning framework to adapt to any task on clients. It achieves this by integrating with a local personalized output layer to accommodate the specific tasks of each client.
- **CreamFL** [16] utilizes a public dataset containing all modalities for knowledge sharing. It captures knowledge from clients through a server–client cross-modality integration strategy based on contrastive learning and regulates training on clients using both inter-modality and intra-modality contrastive learning.
- **CMMC** [12] constructs a cross-modality meta-consensus space to map representations from different modalities into a shared space, enabling cross-modality knowledge sharing. It also designs a universal cross-modality meta-aggregation network to mitigate gradient conflicts across clients, resulting in a more generalized meta-model.

Note that for all **Traditional-FL** models, since they cannot fully adapt to domain-modality-task heterogeneous scenarios, we only perform parameter aggregation among encoder networks (Transformer) of the same modality across different clients. For **Multimodal-FL** models, **CreamFL** achieves cross-client knowledge sharing based on contrastive learning, while **Meta-HAR** and **CMMC** rely on parameter aggregation (meta-learning) for knowledge sharing.

### 5.5. Performance analysis

Table 3 presents a detailed comparison between HeMuGAN and other baseline models across three heterogeneous datasets. Overall, HeMuGAN consistently and significantly outperforms existing Traditional/Multimodal-FL models across all datasets and evaluation metrics. Furthermore, by supporting collaborative learning among clients, HeMuGAN enables each client's local model (based solely on the vanilla Transformer) to outperform the SOTA models specifically designed for each task. It demonstrates that, within our framework, each client can acquire highly valuable external knowledge. In detail, the gains from personalized external knowledge result in an average performance improvement of 8.72% on the Acc-3 and Macro-F1 metrics for the Transformer-based local model on REST14, with improvements of 10.76% on MELD and 9.51% on MOSEI. Similarly, compared to the latest Multimodal-FL model, CMMC, HeMuGAN achieves an average performance improvement of 7.64% on REST14, 9.92% on MELD, and 6.26% on MOSEI. CMMC, which alleviates gradient conflicts through the meta-learning strategy, is the current SOTA

**Table 3**
Comparison of HeMuGAN with other baseline models in distributed collaborative learning across three datasets with domain, modality, and task heterogeneity. The modalities include Uni-modality (text), Bi-modality (text and audio), and Tri-modality (text, audio, and visual). "↓" indicates that a smaller value represents better performance.

| | Models | REST14 Uni-modality<br>Task: SA (%) | | MELD Bi-modality<br>Task: ER (%) | | MOSEI Tri-modality<br>Task: SIR | |
|---|---|---|---|---|---|---|---|
| | | Acc-3 | Macro-F1 | Acc-7 | W-F1 | MAE↓ | Corr |
| Local | SOTA Models [21–23] | 87.31 | 82.27 | 62.84 | 61.12 | 0.5180 | 0.8020 |
| | Transformer [35] | 84.29 | 76.66 | 60.98 | 59.87 | 0.5843 | 0.7703 |
| | Pre-Finetuning | 81.75 | 73.21 | 58.54 | 56.68 | 0.6009 | 0.7427 |
| Traditional FL | Fedavg [48] | 78.32 | 71.16 | 57.22 | 55.93 | 0.6335 | 0.7195 |
| | Fedproto [49] | 79.19 | 72.33 | 58.54 | 56.45 | 0.6238 | 0.7238 |
| | FedGKD [50] | 79.96 | 73.66 | 59.62 | 57.43 | 0.6195 | 0.7308 |
| Multimodal FL | Meta-HAR [51] | 83.56 | 76.69 | 61.45 | 59.36 | 0.5983 | 0.7537 |
| | CreamFL [16] | 85.22 | 76.89 | 61.32 | 60.54 | 0.5852 | 0.7689 |
| | CMMC [12] | 85.59 | 77.03 | 61.34 | 60.43 | 0.5532 | 0.7821 |
| Ours | w/o Cross-Domain transform | 88.29 | 81.72 | 65.39 | 63.94 | 0.5279 | 0.7928 |
| | w/o Cross-Modality reconstruction | 87.30 | 80.13 | 63.25 | 61.77 | 0.5843 | 0.7703 |
| | w/o Cross-Task adaptation | 87.99 | 81.65 | 65.18 | 63.20 | 0.5465 | 0.7823 |
| | HeMuGAN | **89.96** | **84.87** | **67.05** | **66.79** | **0.5078** | **0.8159** |

DCL framework for knowledge sharing via parameter aggregation. The results further indicate that HeMuGAN, by sharing intermediate de-identified representations and employing customized GANs to address domain shifts, modality gaps, and task drifts, presents efficient knowledge-sharing patterns across multiple heterogeneous clients while preserving privacy.

One observation is that, compared to locally trained Transformers, Traditional-FL, which involves collaborative learning across multiple clients, results in worse performance. FedProto and FedGKD can only alleviate the gradient conflicts caused by statistical heterogeneity, but they are no longer capable of addressing the multiple non-statistical heterogeneities. Although Multimodal-FL models are specifically designed to address these issues, the performance improvements for each client remain quite limited. Additionally, such DCL frameworks typically require significant communication overhead, which makes it difficult to achieve a favorable balance between resource consumption and performance improvement. In contrast, HeMuGAN can significantly enhance the performance of each client's tasks, while incurring much lower communication costs (Section 6 for Communication Cost Analysis).

### 5.6. Ablation study

**Functional Ablation** The lower part of Table 3 presents the ablation study results for GANs with different functionalities. Intuitively, removing any component results in a performance decline, underscoring the importance of collaboration among the various GANs in overcoming heterogeneities that hinder knowledge sharing. Specifically, the GAN used for Cross-Domain Transform has the least effect on performance, as its removal does not interfere with the operation of the other components. In contrast, the GAN for Cross-Modality Reconstruction is crucial for completing the missing modality. Its ablation leads to discrepancies in the model input spaces among clients, preventing the execution of subsequent semi-supervised learning and limiting knowledge sharing from clients with more modalities to those with fewer. For the Cross-Task Adaptation GAN, while ablation still permits semi-supervised learning, the lack of optimization for pseudo-labels results in a significant performance decline across all clients.

**Heterogeneity Ablation** Table 4 compares the performance of HeMuGAN with baselines when only domain heterogeneity is present. In this case, the cross-modality reconstruction component of HeMuGAN is not effective, but HeMuGAN still demonstrates a significant performance improvement on REST14. The relatively small performance gain on MOSEI can be attributed to the limited scale of REST14, which results in a lack of external knowledge.

Table 5 compares the performance of HeMuGAN and baseline models under modality heterogeneity only. In this case, the two datasets collaboratively reconstruct the missing modalities, resulting in substantial performance improvement on MOSI. The relatively smaller improvement on MOSEI is attributed to the limited scale of the collaborating dataset.

**Table 4**
Comparison of HeMuGAN with other baseline models across two datasets with only domain heterogeneity.

| Models | REST14 Uni-modality<br>Task: SA (%) | | MOSEI Uni-modality<br>Task: SA (%) | |
|---|---|---|---|---|
| | Acc-3 | Macro-F1 | Acc-2 | F1-Score |
| SOTA Models [21,24] | 87.31 | 82.27 | 82.45 | **82.33** |
| Transformer | 84.29 | 76.66 | 80.21 | 80.15 |
| Pre-Finetuning | 81.75 | 73.21 | 80.27 | 80.19 |
| FedGKD | 81.59 | 73.23 | 73.68 | 73.38 |
| CMMC | 83.38 | 75.34 | 78.92 | 78.76 |
| HeMuGAN (Ours) | **89.97** | **84.09** | **82.47** | 82.29 |

**Table 5**
Comparison of HeMuGAN with other baseline models across two datasets with only modality heterogeneity.

| Models | MOSI Bi-modality<br>Text+Audio Task: SIR | | MOSEI Bi-modality<br>Text+Visual Task: SIR | |
|---|---|---|---|---|
| | MAE↓ | Corr | MAE↓ | Corr |
| SOTA Models [23] | 0.7660 | 0.7460 | 0.6000 | 0.7140 |
| Transformer | 0.7992 | 0.7208 | 0.6513 | 0.6929 |
| Pre-Finetuning | 0.7759 | 0.7382 | 0.6518 | 0.6931 |
| FedGKD | 0.7734 | 0.7294 | 0.6713 | 0.6994 |
| CMMC | 0.7688 | 0.7410 | 0.6554 | 0.6943 |
| HeMuGAN (Ours) | **0.7043** | **0.7921** | **0.5895** | **0.7661** |

Table 6 shows the results under only task heterogeneity. On MOSI, HeMuGAN outperforms the SOTA models. The large-scale, shareable representations of MOSEI further highlight HeMuGAN's ability to acquire knowledge.

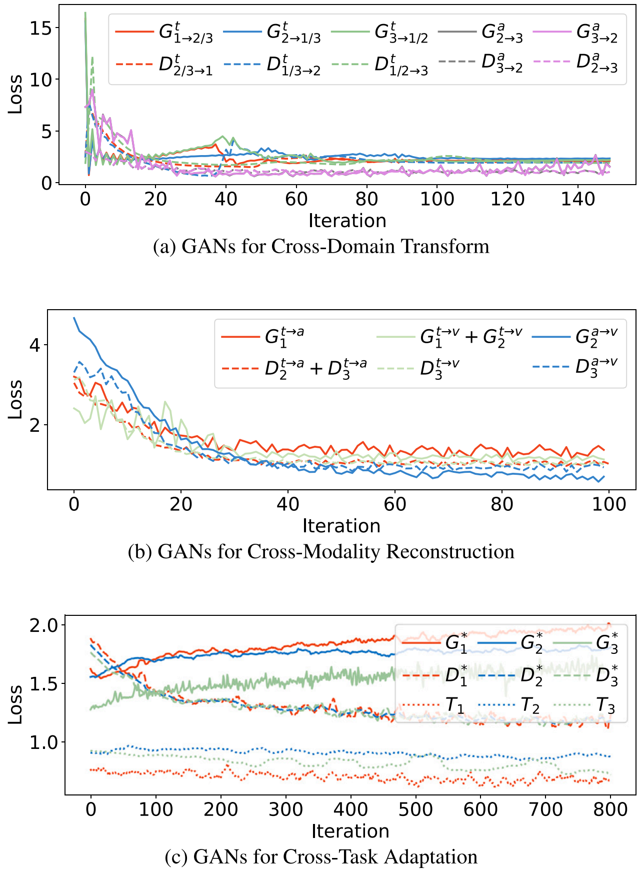
Additionally, HeMuGAN continues to perform well even when there are significant differences in the data scales across clients, without causing any performance degradation. In contrast, for other FL-based baseline models, parameter aggregation under such conditions is a poor choice for large-scale datasets. Gradient conflicts lead to negative transfer, which harms the performance of local models.

### 5.7. Convergence analysis

The adversarial game between the generator and the discriminator can lead to training instability [52]. To illustrate the effectiveness of our customized GANs in mitigating the impact of heterogeneity, Fig. 2 shows the convergence curves of all adversarial components (generator $G$, discriminator $D$, and task predictor $T$) during training in

**Table 6**
Comparison of HeMuGAN with other baseline models across two datasets with only task heterogeneity.

| Models | MOSI Tri-modality<br>Task: SA (%) | | MOSEI Tri-modality<br>Task: SIR | |
|---|---|---|---|---|
| | Acc-2 | F1-Score | MAE↓ | Corr |
| SOTA Models [23] | 86.40 | 86.70 | **0.5180** | **0.8020** |
| Transformer | 82.84 | 82.81 | 0.5843 | 0.7703 |
| Pre-Finetuning | 83.69 | 83.61 | 0.5842 | 0.7727 |
| FedGKD | 83.80 | 83.76 | 0.5880 | 0.7683 |
| CMMC | 84.31 | 84.29 | 0.5889 | 0.7712 |
| HeMuGAN (Ours) | **87.91** | **87.88** | 0.5734 | 0.7814 |



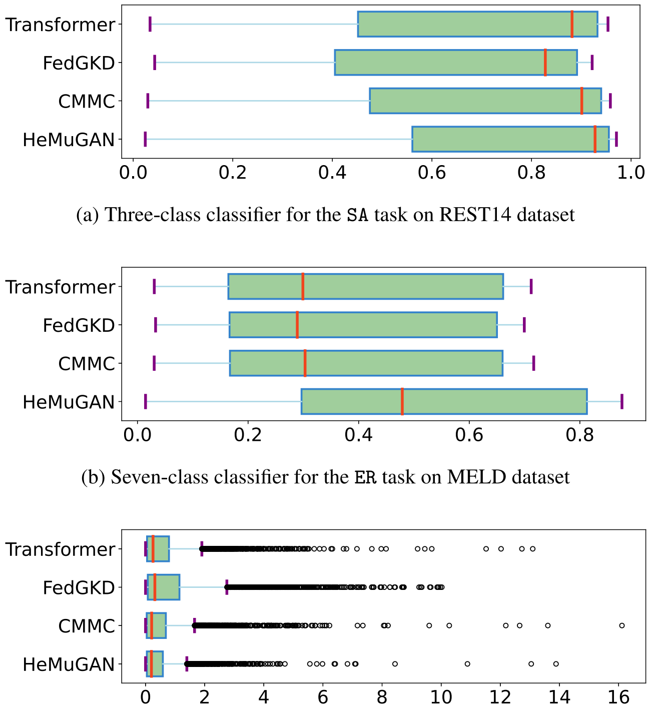
(a) GANs for Cross-Domain Transform

(b) GANs for Cross-Modality Reconstruction

(c) GANs for Cross-Task Adaptation

**Fig. 2.** Visualization of GANs convergence during training on REST14 (1), MELD (2), and MOSEI (3) datasets for HeMuGAN.

HeMuGAN. For GANs applied to cross-domain transform, all generators and discriminators converge to lower loss values, indicating that the generator effectively performs domain transform, making it difficult for robust discriminators to distinguish between external and local representations. For GANs used in cross-modality reconstruction, a similar convergence trend is observed, suggesting that the generators produce "*sufficiently realistic*" reconstructed representations. Notably, for GANs applied to cross-task adaptation, the generator loss exhibits a diverging trend. This occurs because the adversarial objective, in this case, is to improve the task predictor's performance, which requires encouraging the generator to produce "*bad*" samples to increase the predictor's tolerance for erroneous pseudo-labels. This finding aligns with the existing conclusion that effective semi-supervised learning requires a "*bad*" generator [53].



(a) Three-class classifier for the SA task on REST14 dataset

(b) Seven-class classifier for the ER task on MELD dataset

(c) Regression predictor for the SIR task on MOSEI dataset

**Fig. 3.** Visualization of prediction distributions by local task predictors during inference. Specifically, the $x$-axis of (a) and (b) represents the predicted values given by the predictor for the true class of the samples, where higher values indicate better results. In contrast, the $x$-axis of (c) shows the MAE loss between the predicted values and the true values of the samples, where lower values indicate better performance.

## 6. Communication cost analysis

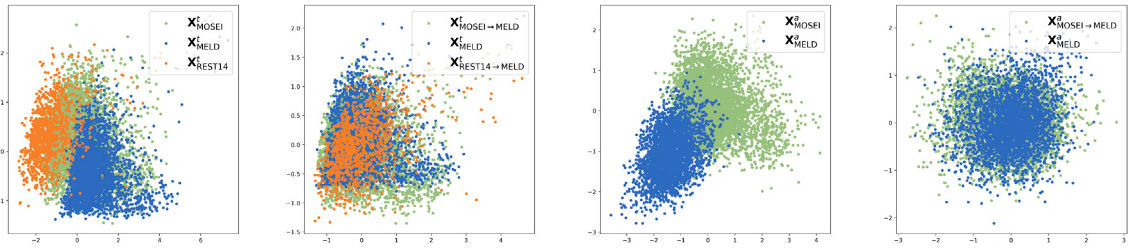
Table 7 compares HeMuGAN with existing FL-based knowledge sharing frameworks in terms of key attributes, along with the analytical expression for the communication cost involved in completing one round of distributed training iteration between two clients. Notably, HeMuGAN avoids gradient conflicts and does not rely on a third-party central server for parameter aggregation. Instead, it facilitates direct communication between the two participating clients, thereby reducing the risk of privacy leakage. Regarding the costs for each communication, HeMuGAN transmits compact representation vectors between clients, which are significantly smaller in size compared to the parameters and gradients of large neural models [54,55]. By contrast, Traditional/Multimodal-FL models require aggregating larger-scale parameters to share more knowledge, leading to higher communication bandwidth demands. Furthermore, between two communication rounds, each client in Traditional/Multimodal-FL frameworks needs to perform one forward and backward propagation on its local full model. By comparison, during the distributed training process of HeMuGAN, parameter updates are limited to the generator and discriminator, which reduces the computational complexity of local training per round, shortens the interval between communication rounds, and ultimately minimizes the overall time consumption. Overall, when performing collaborative learning across multiple heterogeneous clients, HeMuGAN's communication cost and training time remain unaffected by the personalized structure of local models, demonstrating greater efficiency.

## 7. Visualization

To further validate the ability of HeMuGAN to enhance each client's local model by allowing it to autonomously learn the personalized

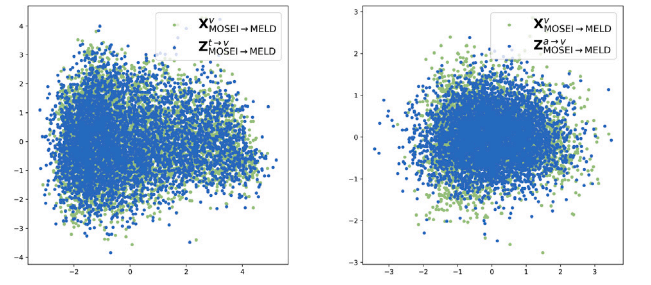
**Table 7**
Comparison of three learning frameworks in key attributes and communication costs, where $\mathbf{W}_i$ and $\mathbf{W}_j$ represent full parameters of the personalized models on Client $i$ and Client $j$, respectively, and $\tau_{\mathbf{W}_i}$ denotes the time required for a single forward and backward propagation on $\mathbf{W}_i$ with a batch size of $B$. Specifically, $|\mathbf{W}_i \cap \mathbf{W}_j|$ is the scale of parameters in the shared sub-block used for parameter aggregation between $i$ and $j$, while $\tau_{agg}$ denotes the time required for parameter aggregation on the server. In addition, $L$ and $d^{\ X/Z}$ refer to the sequence length and dimension of the transmitted representations, respectively. $R$ is the communication rate.

| Terms | Local | Traditional/Multimodal-FL | HeMuGAN (Ours) |
|---|---|---|---|
| Privacy preservation | ✗ | ✓ | ✓ |
| Gradient conflicts | ✗ | ✓ | ✗ |
| Requires a Third-Party central server | ✗ | ✓ | ✗ |
| Parameters per communication | – | $2 \times \lvert\mathbf{W}_i \cap \mathbf{W}_j\rvert$ | $B \times L \times d^{\ X/Z}$ |
| Time interval between two communications | – | $\mathbf{max}(\tau_{\mathbf{W}_i}, \tau_{\mathbf{W}_j})$ | $\tau_{\mathbf{W}_G} + \tau_{\mathbf{W}_D}$ |
| Total time for one distributed training | – | $\mathbf{max}(\tau_{\mathbf{W}_i}, \tau_{\mathbf{W}_j}) + \frac{2\lvert\mathbf{W}_i \cap \mathbf{W}_j\rvert}{R} + \tau_{agg}$ | $\tau_{\mathbf{W}_G} + \frac{B \times L \times d^{\ X/Z}}{R} + \tau_{\mathbf{W}_D}$ |



(a) Text Raw Rep.        (b) Text Transformed Rep.        (c) Audio Raw Rep.        (d) Auido Transformed Rep.

**Fig. 4.** Visualization of representation (Rep.) distribution before and after cross-domain transformation.



(a) Visual Reconstructed from Text        (b) Visual Reconstructed from Audio

**Fig. 5.** Visualization of the distribution between cross-modality reconstructed representations and real representations.

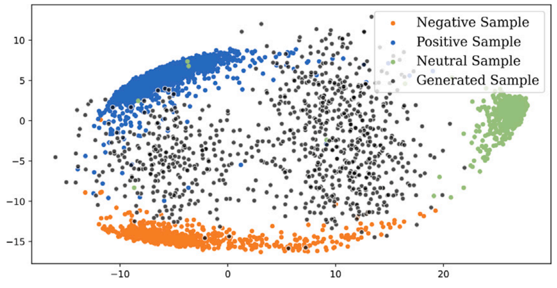


**Fig. 6.** Visualization of the distribution between representations generated by cross-task adaptation GAN and local real samples.

knowledge it requires, we qualitatively visualize the distribution of prediction results of several representative methods for test samples of different datasets in Fig. 3. As shown, HeMuGAN significantly outperforms the comparison models across several key aspects, including the dispersion of prediction results (interquartile range), central tendency (median), skewness (whiskers), and extreme values (outliers).

**Effect of GANs for Cross-Domain Transform** Fig. 4 showcases the capability of our customized GANs to transform the received external representations, aligning their distribution with that of the local representations in the MELD dataset. One clearly observes that, due to domain shifts, there are distribution discrepancies in the common modality representations across different clients, which are effectively mitigated after the transformation. Moreover, this effectiveness is consistently observed in both one-to-one and one-to-many collaborative learning scenarios. The GANs used for cross-domain transformation in HeMuGAN enable each client to seamlessly learn personalized knowledge from both local and external representations, further enhancing the efficiency of collaborative learning.

**Effect of GANs for Cross-Modality Reconstruction** Fig. 5 illustrates the distribution of visual modality representations generated in HeMuGAN, conditioned on text or audio modality representations, alongside the distribution of real visual representations. Evidently, the reconstructed representations and the real representations exhibit highly consistent distributions, indicating that the GANs used for cross-modality reconstruction effectively capture the underlying semantic correlations between different modalities. This enables clients with missing modalities to also benefit from the integration of multimodal information. In addition, a significant advantage of reconstructing missing modalities in HeMuGAN is the elimination of input space discrepancies among clients caused by modality gaps, which further enhances the adaptability of collaborative learning.

**Effect of GANs for Cross-Task Adaptation** Fig. 6 illustrates the distribution of representations for local labeled samples and those generated by the cross-task adaptive GAN in the REST14 dataset. The generated sample points are observed to lie away from the category centers and are distributed near the decision boundaries between categories, aligning with the intended objective [37]. By optimizing the predictor to produce smooth category decisions for these generated samples, it can make consistent decisions for unlabeled samples similarly located near decision boundaries. In semi-supervised learning, this cautious pseudolabel prediction helps mitigate the negative optimization effects caused by incorrect labels guiding the model. HeMuGAN leverages this adversarial semi-supervised learning among local, external, and generated representations, enhancing the robustness of collaborative learning.

**Table 8**
Comparison of inversion results on intermediate representations from different networks under white-box attacks. Higher values indicate greater similarity between the estimated and raw data, while "↓" indicates that lower values correspond to higher similarity. For the audio modality, since the shallow encoder is parameter-free, the white-box attack only attempts to reconstruct the initial MFCCs features extracted by the encoder.

| Models | REST14 text | | | MELD audio | | | MOSEI visual | | |
|---|---|---|---|---|---|---|---|---|---|
| | Rouge-1 | Rouge-2 | Rouge-l | MCD↓ | SNR | DTW↓ | MAE↓ | PSNR | SSIM |
| Single-layer Linear Layer | 34.89 | 6.21 | 27.96 | 1.15 | 18.42 | 0.31 | 0.05 | 61.22 | 0.89 |
| HeMuGAN (Transformer) | 0.17 | 0.01 | 0.43 | 8.58 | 3.27 | 3.51 | 58.43 | 0.77 | 0.12 |

## 8. Privacy analysis

The client's locally private information includes raw data, model architecture, and task labels. HeMuGAN's comprehensive privacy preservation ensures that this information is not directly shared during collaborative learning. However, one concern arises from the potential malicious attacks [56,57] on shared intermediate representations across clients, such as inversion and inference attacks, which could lead to privacy leakage. In this section, we employ advanced white-box attack methods [57] to retrieve the client's local private information by inverting the shareable intermediate representations of the client's model outputs. This method assumes the attacker knows the client's model architecture and constructs an identical threat model. The attacker minimizes the difference between the threat model's output and the shared intermediate representations, applying gradient descent optimization to adjust the estimated inputs of the threat model. This process gradually recovers the client's raw data and steals the parameters of the local model to infer the local task labels. Furthermore, we focus the attack on the shared intermediate representation closest to the raw data, as it is more likely to leak sensitive information that can be used to accurately reconstruct the client's raw data. For HeMuGAN, this corresponds to the output of the cross-domain transform generator from each client.

We also implemented a simple Single Linear Layer on the client side, with its output serving as the shared intermediate representation for attacks, enabling a comparison with HeMuGAN. The inversion results for the raw data from different modalities are shown in Table 8. For the Single Linear Layer, the white-box attack achieved a Rouge-1 score of 34.89, a Rouge-2 score of 6.21, and a Rouge-l score of 27.96 in the text modality, indicating that the threat model effectively recovered textual information. In the audio modality, the attack achieved an MCD score of 1.15, an SNR score of 18.42, and a DTW score of 0.31, demonstrating that the audio MFCC features were also effectively recovered. In the visual modality, the attack achieved an MAE score of 0.05, a PSNR score of 61.22, and an SSIM score of 0.89, showing that the inverted images were highly similar to the originals. These results validate the effectiveness of the white-box attack. However, it struggles to achieve ideal results when applied to HeMuGAN across different modalities. In the text modality, the attack on HeMuGAN resulted in a Rouge-1 score of 0.17, a Rouge-2 score of 0.01, and a Rouge-l score of 0.43, reflecting an average performance reduction of 99.27%. Similar results were observed in the audio and visual modalities, with performance reductions of 91.5% and 99.1% respectively. These results indicate HeMuGAN is effective at resisting attacks, a conclusion consistent with those drawn in previous work [58]. Theoretically proving the security of HeMuGAN's network architecture remains challenging, and we will include this in our future work.

Moreover, since HeMuGAN does not update any parameters based on local task losses during collaborative learning, even if the threat model successfully steals the client's model parameters, local task labels cannot be inferred through reasoning.

## 9. Conclusion

In this paper, we propose a novel DCL framework, HeMuGAN, for knowledge sharing among clients with multiple forms of non-statistical heterogeneities. Unlike existing Multimodal-FL methods, HeMuGAN enables each client to autonomously learn personalized knowledge from the de-identified representations exchanged among clients. Furthermore, we design various customized GANs distributed across clients to eliminate the obstacles posed by domain shifts, modality gaps, and task drifts in knowledge sharing. Experimental results verify that HeMuGAN significantly improves the performance of local models across various heterogeneous scenarios by efficiently acquiring personalized external knowledge while preserving privacy.

## CRediT authorship contribution statement

**Zhuojia Wu:** Writing – original draft, Data curation, Conceptualization. **Qi Zhang:** Writing – review & editing, Funding acquisition, Conceptualization. **Duoqian Miao:** Funding acquisition, Conceptualization. **Xuerong Zhao:** Writing – review & editing. **Guangyin Bao:** Methodology, Investigation. **Liang Hu:** Funding acquisition, Writing – review & editing. **Kun Yi:** Writing – review & editing. **Yu Zhou:** Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This research is supported in part by the National Key Research and Development Program of China (No. 2022YFB3104700), the National Natural Science Foundation of China (No. 62376198, No. 62276190), Shanghai Baiyulan Pujiang Project (No. 08002360429).

## Data availability

The data used in this study are publicly available.

## References

[1] S. Caballe, T. Daradoumis, F. Xhafa, J. Conesa, Enhancing knowledge management in online collaborative learning, Int. J. Softw. Eng. Knowl. Eng. 20 (04) (2010) 485–497.

[2] Z. Wu, Q. Zhang, D. Miao, X. Zhao, K. Shi, Adapting GNNs for document understanding: A flexible framework with multiview global graphs, IEEE Trans. Comput. Soc. Syst. (2024) 1–14, http://dx.doi.org/10.1109/TCSS.2024.3468890.

[3] J. Chen, A. Zhang, On disentanglement of asymmetrical knowledge transfer for modality-task agnostic federated learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 11311–11319.

[4] D.C. Nguyen, M. Ding, P.N. Pathirana, A. Seneviratne, J. Li, H.V. Poor, Federated learning for internet of things: A comprehensive survey, IEEE Commun. Surv. Tutor. 23 (3) (2021) 1622–1658.

[5] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., Advances and open problems in federated learning, Found. Trends® Mach. Learn. 14 (1–2) (2021) 1–210.

[6] Z. Wu, Q. Zhang, D. Miao, K. Yi, W. Fan, L. Hu, HyDiscGAN: A hybrid distributed cGAN for audio-visual privacy preservation in multimodal sentiment analysis, 2024, arXiv preprint arXiv:2404.11938.

[7] T. Nguyen, M.T. Thai, Preserving privacy and security in federated learning, IEEE/ACM Trans. Netw. 32 (1) (2024) 833–843.

[8] C. Thapa, P.C.M. Arachchige, S. Camtepe, L. Sun, Splitfed: When federated learning meets split learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 8485–8493.

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

[10] F. Qi, S. Li, Adaptive hyper-graph aggregation for modality-agnostic federated learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12312–12321.

[11] T. Feng, D. Bose, T. Zhang, R. Hebbar, A. Ramakrishna, R. Gupta, M. Zhang, S. Avestimehr, S. Narayanan, Fedmultimodal: A benchmark for multimodal federated learning, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 4035–4045.

[12] S. Li, F. Qi, Z. Zhang, C. Xu, Cross-modal meta consensus for heterogeneous federated learning, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 975–984.

[13] J. Zhang, Z. Li, B. Li, J. Xu, S. Wu, S. Ding, C. Wu, Federated learning with label distribution skew via logits calibration, in: International Conference on Machine Learning, PMLR, 2022, pp. 26311–26329.

[14] S. Liu, Z. Chen, Y. Liu, Y. Wang, D. Yang, Z. Zhao, Z. Zhou, X. Yi, W. Li, W. Zhang, et al., Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 23436–23446.

[15] X. Yang, B. Xiong, Y. Huang, C. Xu, Cross-modal federated human activity recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2024) 1–18.

[16] Q. Yu, Y. Liu, Y. Wang, K. Xu, J. Liu, Multimodal federated learning via contrastive representation ensemble, 2023, arXiv preprint arXiv:2302.08888.

[17] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.

[18] A. Zadeh, R. Zellers, E. Pincus, L.-P. Morency, Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages, IEEE Intell. Syst. 31 (6) (2016) 82–88.

[19] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2236–2246.

[20] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 527–536.

[21] S. Yin, G. Zhong, TextGT: A double-view graph transformer on text for aspect-based sentiment analysis, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38:17, 2024, pp. 19404–19412.

[22] S. Zou, X. Huang, X. Shen, Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5994–6003.

[23] D. Zong, C. Ding, B. Li, J. Li, K. Zheng, Q. Zhou, AcFormer: An aligned and compact transformer for multimodal sentiment analysis, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 833–842.

[24] J. Huang, Y. Ji, Y. Yang, H.T. Shen, Cross-modality representation interactive learning for multimodal sentiment analysis, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 426–434.

[25] S. Järvelä, D. Gašević, T. Seppänen, M. Pechenizkiy, P.A. Kirschner, Bridging learning sciences, machine learning and affective computing for understanding cognition and affect in collaborative learning, Br. J. Educ. Technol. 51 (6) (2020) 2391–2406.

[26] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2, 2014, pp. 2672–2680.

[27] M. Mirza, S. Osindero, Conditional generative adversarial nets, 2014, arXiv preprint arXiv:1411.1784.

[28] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5907–5915.

[29] J. Gui, Z. Sun, Y. Wen, D. Tao, J. Ye, A review on generative adversarial networks: Algorithms, theory, and applications, IEEE Trans. Knowl. Data Eng. 35 (4) (2021) 3313–3332.

[30] S. Sankaranarayanan, Y. Balaji, C.D. Castillo, R. Chellappa, Generate to adapt: Aligning domains using generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8503–8512.

[31] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.

[32] X. Zhao, Y. Chen, S. Liu, X. Zang, Y. Xiang, B. Tang, TMMDA: a new token mixup multimodal data augmentation for multimodal sentiment analysis, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 1714–1722.

[33] T. Pandeva, M. Schubert, MMGAN: Generative adversarial networks for multi-modal distributions, 2019, ArXiv E-Prints, arXiv–1911.

[34] Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, L. Carin, Adversarial feature matching for text generation, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017, pp. 4006–4015.

[35] A. Vaswani, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017).

[36] M.L. Menéndez, J. Pardo, L. Pardo, M. Pardo, The jensen-shannon divergence, J. Franklin Inst. 334 (2) (1997) 307–318.

[37] J. Dong, T. Lin, Margingan: Adversarial training in semi-supervised learning, Adv. Neural Inf. Process. Syst. 32 (2019).

[38] C. Yang, S. Wang, C. Yang, Y. Li, R. He, J. Zhang, Progressively stacking 2.0: A multi-stage layerwise training method for bert training speedup, 2020, arXiv preprint arXiv:2011.13635.

[39] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, SemEval-2014 task 4: Aspect based sentiment analysis, in: 8th International Workshop on Semantic Evaluation August 23-24, 2014, 2014, pp. 27–35.

[40] T. Shi, S.-L. Huang, MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 14752–14766.

[41] F. Qian, J. Han, Y. He, T. Zheng, G. Zheng, Sentiment knowledge enhanced self-supervised learning for multimodal sentiment analysis, in: Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 12966–12978.

[42] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, O. Nieto, Librosa: Audio and music signal analysis in python, in: SciPy, 2015, pp. 18–24.

[43] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.

[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[45] A. Mollahosseini, B. Hasani, M.H. Mahoor, Affectnet: A database for facial expression, valence, and arousal computing in the wild, IEEE Trans. Affect. Comput. 10 (1) (2017) 18–31.

[46] J. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[47] D.P. Kingma, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[48] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.

[49] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, C. Zhang, Fedproto: Federated prototype learning across heterogeneous clients, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 8432–8440.

[50] D. Yao, W. Pan, Y. Dai, Y. Wan, X. Ding, C. Yu, H. Jin, Z. Xu, L. Sun, FEDGKD: Toward heterogeneous federated learning via global knowledge distillation, IEEE Trans. Comput. 73 (1) (2024) 3–17.

[51] C. Li, D. Niu, B. Jiang, X. Zuo, J. Yang, Meta-har: Federated representation learning for human activity recognition, in: Proceedings of the Web Conference 2021, 2021, pp. 912–922.

[52] A. Radford, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015, arXiv preprint arXiv:1511.06434.

[53] Z. Dai, Z. Yang, F. Yang, W.W. Cohen, R. Salakhutdinov, Good semi-supervised learning that requires a bad GAN, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6513–6523.

[54] A. Singh, P. Vepakomma, O. Gupta, R. Raskar, Detailed comparison of communication efficiency of split learning and federated learning, 2019, arXiv preprint arXiv:1909.09145.

[55] P. Vepakomma, T. Swedish, R. Raskar, O. Gupta, A. Dubey, No peek: A survey of private distributed deep learning, 2018, arXiv preprint arXiv:1812.03288.

[56] D. Pasquini, G. Ateniese, M. Bernaschi, Unleashing the tiger: Inference attacks on split learning, in: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 2113–2129.

[57] E. Erdoğan, A. Küpçü, A.E. Çiçek, Unsplit: Data-oblivious model inversion, model stealing, and label inference attacks against split learning, in: Proceedings of the 21st Workshop on Privacy in the Electronic Society, 2022, pp. 115–124.

[58] J.-Y. Zheng, H. Zhang, L. Wang, W. Qiu, H.-W. Zheng, Z.-M. Zheng, Safely learning with private data: A federated learning framework for large language model, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 5293–5306.