Bridging Gaps: Federated Multi-View Clustering in Heterogeneous Hybrid Views

Xinyue Chen[1] **Yazhou Ren**[1,2,*] **Jie Xu**[1]
Fangfei Lin[1] **Xiaorong Pu**[1,2] **Yang Yang**[1]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, China; [2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, China

Abstract

Recently, federated multi-view clustering (FedMVC) has emerged to explore cluster structures in multi-view data distributed on multiple clients. Many existing approaches tend to assume that clients are isomorphic and all of them belong to either single-view clients or multi-view clients. While these methods have succeeded, they may encounter challenges in practical FedMVC scenarios involving heterogeneous hybrid views, where a mixture of single-view and multi-view clients exhibit varying degrees of heterogeneity. In this paper, we propose a novel FedMVC framework, which concurrently addresses two challenges associated with heterogeneous hybrid views, i.e., client gap and view gap. To address the client gap, we design a local-synergistic contrastive learning approach that helps single-view clients and multi-view clients achieve consistency for mitigating heterogeneity among all clients. To address the view gap, we develop a global-specific weighting aggregation method, which encourages global models to learn complementary features from hybrid views. The interplay between local-synergistic contrastive learning and global-specific weighting aggregation mutually enhances the exploration of the data cluster structures distributed on multiple clients. Theoretical analysis and extensive experiments demonstrate that our method can handle the heterogeneous hybrid views in FedMVC and outperforms state-of-the-art methods. The code is available at `https://github.com/5Martina5/FMCSC`.

1 Introduction

Recent advancements in sensors and the internet have allowed many distributed clients to collect unlabeled data from multiple views/modalities[10, 17, 25]. Utilizing these unlabeled data while considering the need for data privacy among clients has given rise to an emerging field of federated multi-view clustering (FedMVC) [8], which enables multiple clients to collaboratively train consistent clustering models without exposing private data. The clustering models across distributed clients can be applied in many applications (e.g., recommendation [15] and medicine [5]) and thus FedMVC attracts increasing research interest [14, 26, 35].

Existing FedMVC methods tend to assume that clients are isomorphic and all of them belong to either single-view clients or multi-view clients. For instance, FedDMVC [8] assumes that a dataset with V views is distributed across V single-view clients, each having the same set of samples. FedMVFPC [14] assumes that the data are distributed among multiple multi-view clients, with each client having V views and non-overlapping samples among clients. Despite the achieved success, they may encounter challenges when handling some practical FedMVC scenarios involving heterogeneous

*Corresponding author (yazhou.ren@uestc.edu.cn).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-1168

hybrid views, where different clients are heterogeneous such that single-view clients and multi-view clients are hybrid.

This scenario is prevalent in real world situations [6, 47], for example, hospitals in metropolitan areas employ CT, X-ray, and EHR for disease detection, whereas remote areas usually rely on a single detection method. Similarly, smartphones can simultaneously capture audio and images, but recording devices are limited to collecting audio data only.

The presence of heterogeneous hybrid views might limit the applicability of previous FedMVC methods. We decompose these challenges into two issues. (a) *Client gap*. Multi-view clients collect multiple views that have the opportunity to learn comprehensive cluster partitions by leveraging multi-view information. Conversely, single-view clients only own single views and easily obtain biased cluster partitions if the single view only contains marginal information. (b) *View gap*. Hybrid views collected by different clients have quality differences (e.g., images might contain richer visual information than texts, but texts could have semantic information), and extracting complementary information from different views across all clients is not trivial.

In this paper, we introduce a novel FedMVC method, namely Federated Multi-view Clustering via Synergistic Contrast (FMCSC), which can simultaneously leverage the single-view and multi-view data across heterogeneous clients to mine clustering structures from hybrid views. Figure 1 shows an overview of our FMCSC framework. First, we note that lacking unified information supervision, naive aggregation of local models may lead to model misalignment. Therefore, we propose cross-client consensus pre-training to align the local models on all clients to avoid their misalignment. Second, to address the client gap, we design local-synergistic contrastive learning that helps to mitigate the heterogeneity between single-view clients and multi-view clients. In particular, we leverage feature-level and model-level contrastive learning to align multi-view clients and single-view clients, respectively. Third, to tackle the view gap, we develop the global-specific weighting aggregation which encourages global models to learn robust features from hybrid views, further exploring complementary cluster structures. The local-synergistic contrastive learning and global-specific weighting aggregation promote each other to explore the data cluster structures distributed on multiple clients. Overall, FMCSC effectively facilitates all clients in bridging the client gap and view gap within the heterogeneous hybrid views through theoretical and experimental analysis.

Our main contributions are as follows:

- We propose a novel FedMVC method that can handle the heterogeneous hybrid views and explain the success mechanism of the proposed method through theoretical analysis from the perspective of bridging client and view gaps.
- We design local-synergistic contrastive learning and global-specific weighting aggregation, using mutual information as a bridge to connect local and global models, together to explore the cluster structures in multi-view data distributed on different clients.
- Theoretical and experimental analyses verify the effectiveness of FMCSC, which shows excellent clustering performance under various federated learning scenarios.

2 Related Work

Federated multi-view clustering (FedMVC) has emerged to explore cluster structures in multi-view data distributed on multiple clients. Existing FedMVC methods can be classified into two categories based on the partitioning of multi-view data among clients. (1) Vertical FedMVC assumes that a dataset with V views is distributed across V single-view clients, each having the same set of samples. Robust federated multi-view learning (FedMVL) [16] addresses high communication costs, fault tolerance, and stragglers. Federated deep multi-view clustering (FedDMVC) [8] focuses on addressing the challenges of feature heterogeneity and incompleteness. Existing vertical FedMVC methods still rely on the idealistic assumption that different views of the same sample can be aligned across clients, which warrants further investigation. (2) Horizontal FedMVC assumes that the data are distributed among multiple multi-view clients, with each client having V views and non-overlapping samples among clients. Federated multi-view fuzzy c-means consensus prototypes clustering (FedMVFPC) [14] utilizes federated learning mechanisms to perform fuzzy c-means clustering on multi-view data. Horizontal federated multi-view learning (H-FedMV) [5] aims to improve the local disease prediction performance by sharing training models among clients. Although existing approaches have been

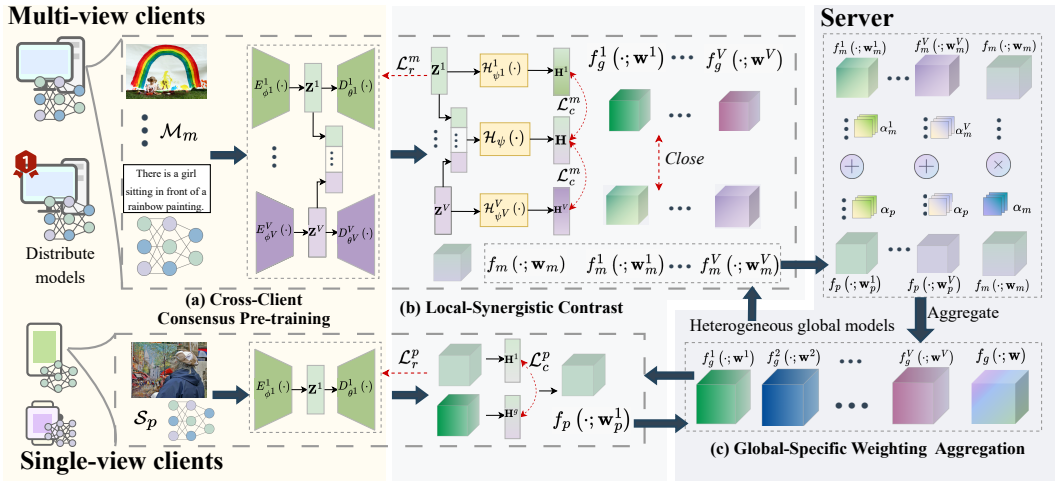


Figure 1: The framework of FMCSC. Initially, each client conducts cross-client consensus pre-training to alleviate model misalignment (Section 3.2). Then, all clients begin training using the designed local-synergistic contrast (Section 3.3) and upload their local models to the server. The server performs global-specific weighting aggregation and distributes multiple heterogeneous global models to all clients (Section 3.4). Finally, leveraging global models received from the server, clients discover complementary cluster structures across all clients.

successful, they may encounter challenges in practical FedMVC scenarios with heterogeneous hybrid views. Specifically, heterogeneity refers to the differences among clients, where single-view and multi-view clients coexist. The hybrid views indicate the uncertainty in the number and quality of views involved in training. Our proposed FMCSC is a variant of horizontal FedMVC, and can handle such scenarios by bridging the client gap and the view gap among clients.

3 Methodology

The key goal of FMCSC is to bridge client and view gaps. On the one hand, multi-view clients have the opportunity to learn comprehensive cluster partitions by leveraging multi-view information. Single-view clients aim to bridge the client gap between themselves and multi-view clients, thereby avoiding obtaining biased clustering partitions. On the other hand, considering the inherent discrepancies in data quality among different views, our objective is to extract complementary information from hybrid views across all clients.

3.1 Problem Definition

We consider a heterogeneous federated learning setting where M multi-view clients and S single-view clients collaborate to train and mine complementary clustering structures using multiple heterogeneous global models $\left\{f_g^1\left(\cdot;\mathbf{w}^1\right),\ldots,f_g^V\left(\cdot;\mathbf{w}^V\right),f_g\left(\cdot;\mathbf{w}\right)\right\}$. Here, $f_g^v\left(\cdot;\mathbf{w}^v\right)$ represents the global model capable of handling the v-th view type, and $f_g\left(\cdot;\mathbf{w}\right)$ represents the global model capable of handling multi-view data. Each single-view client $p\in[S]$ has its private dataset $\mathcal{S}_p=\{\mathbf{x}_i^v\}_{i=1}^{|\mathcal{S}_p|}$, where $\mathbf{x}_i^v$ represents the i-th sample of the p-th single-view client collected from the v-th view type. It adopts a small model $f_p\left(\cdot;\mathbf{w}_p^v\right):\mathbb{R}^{D_v}\to\mathbb{R}^d$ based on its local view types. The multi-view client $m\in[M]$ has its local dataset $\mathcal{M}_m=\left\{\left(\mathbf{x}_i^1,\mathbf{x}_i^2,\ldots,\mathbf{x}_i^V\right)\right\}_{i=1}^{|\mathcal{M}_m|}$, where $\left(\mathbf{x}_i^1,\mathbf{x}_i^2,\ldots,\mathbf{x}_i^V\right)$ represents the i-th sample of the m-th multi-view client collected from V different view types. It trains a small model $f_m\left(\cdot;\mathbf{w}_m\right):\mathbb{R}^{\sum_{v=1}^V D_v}\to\mathbb{R}^d$ based on its local data. For simplicity, we assume that the output features of all models have the same dimension d.

https://doi.org/10.52202/079017-1168

3.2 Cross-Client Consensus Pre-training

Multi-view datasets often contain redundancy and random noise. Current mainstream methods employ self-supervised autoencoder models, such as autoencoder (AE) [13] and variational autoencoder (VAE) [20] to learn high-level features from raw data. In FMCSC, we employ an encoder-decoder pair, denoted $E^v_{\phi^v}(\cdot)$ and $D^v_{\theta^v}(\cdot)$ with learnable parameters ϕ^v and θ^v, for each view in each client. For multi-view client m, we define $\mathbf{z}_i^v = E^v_{\phi^v}(\mathbf{x}_i^v) \in \mathbb{R}^{d_v}$ as the d_v-dimensional latent feature of the i-th sample, and the output of the autoencoder is $\hat{\mathbf{x}}_i^v = D^v_{\theta^v}(\mathbf{z}_i^v) \in \mathbb{R}^{D_v}$. We calculate the reconstruction loss between the input $\mathbf{x}_i^v$ and the output $\hat{\mathbf{x}}_i^v$ for all samples in this client. Additionally, the local model of this client consists of V encoder-decoder pairs, which can be pre-trained by minimizing the reconstruction objective:

$$\mathcal{L}_r^m = \frac{1}{|\mathcal{M}_m|} \sum_{v=1}^{V} \sum_{i=1}^{|\mathcal{M}_m|} \left\| \mathbf{x}_i^v - D^v_{\theta^v}\left(E^v_{\phi^v}(\mathbf{x}_i^v)\right) \right\|_2^2. \tag{1}$$

Similarly, for single-view client p that contains a type of view data locally, it is sufficient to construct an encoder-decoder pair. Pre-training can be performed using the same objective as in Eq. (1):

$$\mathcal{L}_r^p = \frac{1}{|\mathcal{S}_p|} \sum_{i=1}^{|\mathcal{S}_p|} \left\| \mathbf{x}_i^v - D^v_{\theta^v}\left(E^v_{\phi^v}(\mathbf{x}_i^v)\right) \right\|_2^2. \tag{2}$$

Key Observations: In federated learning, diverse local data distributions frequently lead to model drift, causing slow and unstable convergence [19, 47]. In FMCSC, single-view clients have never encountered data from other view types, thus intensifying the issue of model drift due to heterogeneous hybrid views. Moreover, the absence of uniformly labeled data across all clients allows the reconstruction objective of autoencoders to optimize from multiple different directions. In feature space, this problem manifests itself as angular deviations among features [48], and from the perspective of model aggregation, it manifests itself as model misalignment. In other words, direct aggregation of models can blur the feature distinctions captured by local models, leading to inseparability among features, as shown in Figure 3.

A direct strategy to alleviate model misalignment is through alignment. Based on this naive idea, we propose that the multi-view client that finishes training first should distribute its network parameters $\left\{E^1_{\phi^1}(\cdot), \ldots, E^V_{\phi^V}(\cdot)\right\}$ and $\left\{D^1_{\theta^1}(\cdot), \ldots, D^V_{\theta^V}(\cdot)\right\}$ to the remaining clients. Each client then performs pre-training based on this model, thereby alleviating the model misalignment caused by unsupervised training. Notably, as pre-training solely involves training the autoencoder, the construction of local models is inherently dependent on the view type. Therefore, single-view clients can still refer to the network parameters of multi-view clients. This process facilitates consensus pre-training among the clients, ultimately leading to the uploading of pre-trained model parameters to the server. The server initializes global models based on the models uploaded by clients.

3.3 Local-Synergistic Contrast

During pre-training, the features extracted through the reconstruction objective usually contain both common semantics and view-private information. The latter is often meaningless or even misleading, leading to poor clustering effectiveness. To mitigate the adverse effects of view-private information, each client also needs to design a consistent objective during training to learn common semantics.

For multi-view client m, it possesses information from multiple views. Inspired by many previous works of MVC [40, 42, 44, 45], we employ feature contrastive learning to achieve consistency objectives. Considering the conflict between consistency and reconstruction objectives, we opt to operate in distinct feature spaces. We refer to the features obtained through the autoencoders $\left\{\left(\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^V\right)\right\}_{i=1}^{|\mathcal{M}_m|}$ as low-level features. Moreover, we stack V non-linear mappings $\left\{\mathcal{H}^1\left(\mathbf{Z}^1; \Psi^1\right), \ldots, \mathcal{H}^V\left(\mathbf{Z}^V; \Psi^V\right)\right\}$ to obtain high-level features $\left\{\left(\mathbf{h}_i^1, \mathbf{h}_i^2, \ldots, \mathbf{h}_i^V\right)\right\}_{i=1}^{|\mathcal{M}_m|}$. Additionally, a non-linear mapping $\mathcal{H}(\mathbf{Z}; \Psi) : \mathbb{R}^{\sum_{v=1}^{V} d_v} \to \mathbb{R}^d$ is constructed by:

$$\mathbf{H} = \mathcal{H}(\mathbf{Z}; \Psi) = \mathcal{H}\left(\left[\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^V\right]; \Psi\right), \tag{3}$$

where $\{\mathbf{h}_i\}_{i=1}^{|\mathcal{M}_m|} = \mathbf{H} \in \mathbb{R}^{|\mathcal{M}_m| \times d}$, and $\mathbf{Z} \in \mathbb{R}^{|\mathcal{M}_m| \times \sum_{v=1}^{V} d_v}$. We aim to preserve the representative capacity of low-level features to prevent model collapse while learning the common semantics $\mathbf{H}$ among views in the high-level feature space.

In the high-level feature space, the common semantics $\mathbf{H}$ are learned from all views and should be very similar to the common semantics learned from individual views. Based on this, we define $\left\{\mathbf{h}_i, \mathbf{h}_j^v\right\}_{j=i}^{v=1,\ldots,V}$ as V positive feature pairs, and the remaining $\left\{\mathbf{h}_i, \mathbf{h}_j^v\right\}_{j\neq i}^{v=1,\ldots,V}$ are $V(|\mathcal{M}_m| - 1)$ negative feature pairs. Then, we use cosine similarity to measure the similarity of feature pairs:

$$sim\left(\mathbf{h}_i, \mathbf{h}_j^v\right) = \frac{\left\langle \mathbf{h}_i, \mathbf{h}_j^v \right\rangle}{\|\mathbf{h}_i\| \left\|\mathbf{h}_j^v\right\|}, \tag{4}$$

where $\langle \cdot, \cdot \rangle$ is dot product operator. We introduce the temperature parameter τ_m to moderate the effect of similarity. Subsequently, the feature contrastive loss is formulated as:

$$\mathcal{L}_c^m = -\frac{1}{|\mathcal{M}_m|} \sum_{v=1}^{V} \sum_{i=1}^{|\mathcal{M}_m|} \log \frac{e^{sim(\mathbf{h}_i, \mathbf{h}_i^v)/\tau_m}}{\sum_{j\neq i} e^{sim\left(\mathbf{h}_i, \mathbf{h}_j^v\right)/\tau_m}}. \tag{5}$$

Moreover, multi-view clients aim to assist single-view clients in bridging client gaps and discarding view-private information detrimental to clustering. They optimize multiple heterogeneous global models using local data, ensuring that even the global model designed for single-view processing acquires generalized common semantics. Such common semantics are also advantageous for uncovering complementary clustering structures across clients. Concretely,

$$\min_{\{\mathbf{w}_m^v\}_{v=1}^{V}} \sum_{v=1}^{V} \left\| f_m^v\left(\cdot; \mathbf{w}_m^v\right) - f_m\left(\cdot; \mathbf{w}_m\right) \right\|_2^2, \tag{6}$$

where $f_m^v\left(\cdot; \mathbf{w}_m^v\right)$ represents the local model that can handle the v-th type of view, which is initialized by the global model $f_g^v\left(\cdot; \mathbf{w}^v\right)$. $f_m\left(\cdot; \mathbf{w}_m\right)$ represents the local model of multi-view client m, where multi-view clients possess data of all view types.

For single-view client p, where each sample only has a single view, we design a model contrastive learning to achieve consistency objectives. Specifically, client p contains data of the v-th view type $\{\mathbf{x}_i^v\}_{i=1}^{|\mathcal{S}_p|}$, with its local model as $f_p\left(\cdot; \mathbf{w}^v\right)$. To explore common semantics, we adopt the same approach as multi-view clients by constructing two non-linear mappings $\mathcal{H}^v\left(\mathbf{Z}^v; \Psi^v\right)$ and $\mathcal{H}\left(\mathbf{Z}; \Psi\right)$ to obtain high-level features $\mathbf{H}^v$ and common semantics $\mathbf{H}$. The global model $f_g\left(\cdot; \mathbf{w}^v\right)$ after aggregation further enhances its ability to learn generalized common semantics. To encourage the local model of client p to approach the more generalized global model, we formulate the model contrastive loss:

$$\mathcal{L}_c^p = -\frac{1}{|\mathcal{S}_p|} \sum_{i=1}^{|\mathcal{S}_p|} \log \frac{e^{sim\left(\mathbf{h}_i, \mathbf{h}_i^g\right)/\tau_p}}{e^{sim\left(\mathbf{h}_i, \mathbf{h}_i^g\right)/\tau_p} + e^{sim\left(\mathbf{h}_i, \mathbf{z}_i^v\right)/\tau_p}}, \tag{7}$$

where $\{\mathbf{h}_i^g\}_{i=1}^{|\mathcal{S}_p|}$ represents the output of the local data after being processed by the global model, and τ_p denotes the temperature parameter. The significance of Eq. (7) lies in treating $\{\mathbf{h}_i, \mathbf{h}_i^g\}_{i=1}^{|\mathcal{S}_p|}$ as positive pairs and $\{\mathbf{h}_i, \mathbf{z}_i^v\}_{i=1}^{|\mathcal{S}_p|}$ as negative pairs. This allows the local model of client p to converge towards the global model while amplifying the differences between the reconstruction and consistency objectives in the local model.

During training, the respective total losses for multi-view client m and single-view client p are:

$$\mathcal{L}^m = \mathcal{L}_r^m + \mathcal{L}_c^m, \quad \mathcal{L}^p = \mathcal{L}_r^p + \mathcal{L}_c^p. \tag{8}$$

In the optimization of FMCSC, $\mathcal{L}_r^m$ and $\mathcal{L}_r^p$ are utilized as reconstruction losses to learn representations for each view individually. Meanwhile, $\mathcal{L}_c^m$ and $\mathcal{L}_c^p$ are employed to discover common semantics across views, facilitating the exploration of complementary clustering structures across clients.

3.4 Global-Specific Weighting Aggregation

To bridge the view gap and aggregate heterogeneous models, we design a weighted specific aggregation strategy on the server, yielding multiple heterogeneous global models.

https://doi.org/10.52202/079017-1168

Theorem 1. *Assume $\delta_m, \delta_p \in (0,1)$ such that $p(\mathbf{h}_i^v \mid \mathbf{h}_i) > \delta_m$, $i = 1, 2, \cdots, |\mathcal{M}_m|$ and $p(\mathbf{h}_i^g \mid \mathbf{h}_i) = p(\mathbf{z}_i^v \mid \mathbf{h}_i) > \delta_p$, $i = 1, 2, \cdots, |\mathcal{S}_p|$ hold. The following inequality establishes the relationship between the consistency objectives and the mutual information of the multi-view client m and single-view client p, respectively:*

$$\begin{aligned} &\sum_{v=1}^{V} I(\mathbf{H}, \mathbf{H}^v) \geq V \log |\mathcal{M}_m| - \delta_m \mathcal{L}_c^m, \\ &I(\mathbf{H}, \mathbf{H}^g) - I(\mathbf{H}, \mathbf{Z}^v) \leq -2\delta_p \mathcal{L}_c^p. \end{aligned} \tag{9}$$

Proofs of all the theorems in this paper are provided in Appendix C due to space limit. Theorem 1, Eq. (5), and Eq. (7) indicate that minimizing contrastive loss $\mathcal{L}_c^m$ and $\mathcal{L}_c^p$ are equal to maximizing mutual information. Such connection has also been discussed in [42, 51].

Through theoretical analysis, we use mutual information $\sum_{v=1}^{V} I(\mathbf{H}, \mathbf{H}^v)$ and $I(\mathbf{H}, \mathbf{H}^g) - I(\mathbf{H}, \mathbf{Z}^v)$ as weights to evaluate the quality of the models from multi-view clients and single-view clients, respectively. A higher level of mutual information indicates better model quality, leading to higher weights during aggregation.

Considering the heterogeneity of the models, we aggregate client models with the same architecture on the server, referred to as specific aggregation. Specifically,

$$\begin{aligned} f_g(\cdot; \mathbf{w}) &= \sum_{m=1}^{M} \alpha_m f_m(\cdot; \mathbf{w}_m), \\ f_g^v(\cdot; \mathbf{w}^v) &= \sum_{m=1}^{M} \alpha_m^v f_m^v(\cdot; \mathbf{w}_m^v) + \sum_{p=1}^{S} \alpha_p f_p\left(\cdot; \mathbf{w}_p^v\right), \end{aligned} \tag{10}$$

where $v = 1, 2, \cdots, V$, α_m and α_p represent the weights for model aggregation of multi-view client m and single-view client p respectively. In this scenario, there are a total of M multi-view clients and S single-view clients, resulting in $(V+1)$ heterogeneous global models.

For global models that handle multiple view types simultaneously, the expected risk is defined as $\mathcal{L}_M(f)$ and optimized by minimizing the empirical risk $\widehat{\mathcal{L}}_M(f)$. Similarly, for global models dealing with processing a single view type, such as the v-th view type, the expected and empirical risks are defined as $\mathcal{L}_{S^v}(f)$ and $\widehat{\mathcal{L}}_{S^v}(f)$ respectively. Inspired by previous works [23, 37] on the generalization bound of clustering approaches, we obtain the following theorem by analyzing the generalization bound of the proposed FMCSC method.

Theorem 2. *Suppose that for any $\mathbf{x} \in \mathcal{X}$ and $f \in \mathcal{F}$, there exists $D < \infty$ such that $\|\mathbf{x}\|, \|f_x(\mathbf{x})\|, \|f_z^v(\mathbf{x})\|, \|f_h^v(\mathbf{x})\|, \left\|f_h^0(\mathbf{x})\right\| \in [0, D]$ hold. With probability $1-\delta$ for any $f \in \mathcal{F}$, the following inequality holds*

$$\begin{aligned} \mathcal{L}_M(f) \leq& \widehat{\mathcal{L}}_M(f) + \frac{12VD^2}{\sqrt{M|\mathcal{M}_m|}} + 9VD^2\sqrt{\frac{\log\frac{1}{\delta}}{2M|\mathcal{M}_m|}}, \\ L_{S^v}(f) \leq& \widehat{L}_{S^v}(f) + \frac{10D^2}{\sqrt{S^v|\mathcal{S}_p|}} + 8D^2\sqrt{\frac{\log\frac{1}{\delta}}{2S^v|\mathcal{S}_p|}} \\ &+ \sqrt{\frac{4}{M|\mathcal{M}_m|}\left(d\log\frac{2eM|\mathcal{M}_m|}{d} + \log\frac{4}{\delta}\right)} + d_{\mathcal{F}}\left(\tilde{\mathcal{D}}_v, \tilde{\mathcal{D}}\right) + \lambda_v, \end{aligned} \tag{11}$$

where M is the number of multi-view participating clients, S^v is the number of single-view clients with v-th view type that participated in the training, $|\mathcal{M}_m|$ and $|\mathcal{S}_p|$ are the number of samples in each multi-view client and single-view client, respectively. $d_{\mathcal{F}}\left(\tilde{\mathcal{D}}_v, \tilde{\mathcal{D}}\right)$ measures the difference between data from the v-th view distribution $\tilde{\mathcal{D}}_v$ and the multi-view data distribution $\tilde{\mathcal{D}}$.

Three key implications can be derived from Theorem 2: i) For global models capable of handling multi-view data, having more samples from multi-view clients, such as increasing the number of samples per client or adding multi-view participating clients, contributes to the improvement of

generalization performance. ii) For global models dealing with single-view data, having more samples from both multi-view and single-view clients enhances their generalization performance, which is a reflection of multi-view clients helping single-view clients to bridge the client gap. iii) Although the view gap is mitigated, the high dissimilarity among views still leads to high distribution divergence $d_{\mathcal{F}}\left(\tilde{\mathcal{D}}_v, \tilde{\mathcal{D}}\right)$, which impairs the quality of the global model.

Finally, each client applies the K-means [30] on the common semantics $\mathbf{H}$ obtained from the corresponding global model to calculate the cluster centroids and obtain their local clustering results. For example, for multi-view client m, letting $\{\mathbf{c}_j\}_{j=1}^{K}$ denote the K cluster centroids, we have:

$$\min_{\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_K} \sum_{i=1}^{|\mathcal{M}_m|} \sum_{j=1}^{K} \|\mathbf{h}_i - \mathbf{c}_j\|^2 . \tag{12}$$

The clustering result for the i-th sample is $y_i = \arg\min_j \|\mathbf{h}_i - \mathbf{c}_j\|^2$. By concatenating clustering results from all clients, we obtain the overall clustering results.

4 Experiments

4.1 Experimental Settings

Datasets. Our experiments are carried out on four multi-view datasets. Specifically, **MNIST-USPS** [34] comprises 5000 samples collected from two handwritten digital image datasets, which are considered as two views. **BDGP** [4] consists of 2500 samples across 5 drosophila categories, with each sample having textual and visual views. **Multi-Fashion** [41] contains images from 10 categories, where we treat three different styles of one object as three views, resulting in 10000 samples. **NUSWIDE** [9] consists of 5000 samples obtained from web images with 5 views. Considering the sample quantities, we allocate BDGP to 12 clients, MNIST-USPS and NUSWIDE to 24 clients respectively, and Multi-Fashion to 48 clients to simulate the federated learning settings.

Comparison Methods. We select 9 state-of-the-art methods, including HCP-IMSC [24], IMVC-CBG [39], DSIMVC [37], LSIMVC [27], ProImp [22], JPLTD [29], CPSPAN [18], FedDMVC [8] and FCUIF [36]. Among them, apart from FedDMVC and FCUIF, which are FedMVC methods, all the other comparison methods are centralized incomplete multi-view clustering methods. To ensure fair comparisons, we concatenate the data distributed among the clients and use them as the input for centralized methods. Among these, the data from multi-view clients can be regarded as complete data, while the data from single-view clients can be considered as missing data.

Implementation Details. For an encoder-decoder pair, the encoder structure is Input- $\mathrm{Fc}_{500} - \mathrm{Fc}_{500} - \mathrm{Fc}_{2000} - \mathrm{Fc}_{20}$, and the decoder is symmetric with the encoder. Also, we set temperature parameters $\tau_m = \tau_p = 0.5$ and use the batch size of 256. The output dimension d is set to 20 for all local and global models and communication rounds R is set to 5. All experiments in the paper involving FMCSC that are not mentioned are performed when M/S = 1:1.

4.2 Results and Analysis

Clustering Results. Table 1 presents a quantitative comparison under various heterogeneous hybrid view scenarios. Each experiment is independently conducted five times, reporting average values and standard deviations. We construct different scenarios by adjusting the ratio of multi-view clients to single-view clients. Remarkably, FMCSC achieves exceptional performance in various heterogeneous hybrid view scenarios, surpassing recent methods. This indicates our ability to achieve satisfactory clustering performance while preserving data privacy. Furthermore, with the increasing proportion of single-view clients, all methods exhibit varying degrees of performance decline, aligning with common expectations. Even when the number of single-view clients is twice that of multi-view clients, FMCSC still achieves superior clustering performance, which is encouraging.

Ablation Studies. Components A and D represent the consensus pre-training process and weighted aggregation, respectively. Component B indicates that multi-view clients bring the global models closer as Eq. (6). Component C indicates that model comparison in single-view clients as Eq. (7). Table 2 shows that the impact of component A on clustering performance is dependent on the dataset. Additionally, note that the results in Item-1 are obtained after running four times

 https://doi.org/10.52202/079017-1168

Table 1: Clustering results (mean±std%) of all methods on four datasets. The best and second best results are denoted in **bold** and underline.

	Multi- / Single- clients	M/S = 2:1			M/S = 1:1			M/S = 1:2		
	Evaluation metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
MNIST-USPS	HCP-IMSC (2022) [24]	80.2±0.0	74.8±0.0	69.8±0.1	79.0±0.2	71.6±0.2	66.9±0.2	76.2±0.1	71.0±0.1	63.6±0.1
	IMVC-CBG (2022) [39]	46.6±0.1	40.3±0.2	22.2±0.4	38.9±0.1	35.3±0.3	14.9±0.2	37.3±0.6	31.7±0.4	10.6±0.2
	DSIMVC (2022) [37]	55.1±0.1	27.6±0.1	25.0±0.1	54.5±0.1	26.7±0.1	24.4±0.2	54.1±0.2	26.6±0.2	24.0±0.3
	LSIMVC (2022) [27]	59.3±0.2	55.2±0.9	38.2±1.7	52.7±0.4	46.5±0.2	25.8±0.5	41.8±0.4	37.1±0.4	14.2±0.5
	ProImp (2023) [22]	91.2±0.9	84.4±1.0	80.8±2.1	87.3±0.6	77.9±0.7	73.1±1.0	84.8±0.9	75.9±0.9	66.6±1.8
	JPLTD (2023) [29]	40.7±0.1	22.6±0.1	17.3±0.0	40.0±0.2	19.8±0.1	15.2±0.1	32.3±0.1	13.7±0.1	10.1±0.1
	CPSPAN (2023) [18]	79.5±2.5	77.3±2.0	70.7±2.2	76.5±2.7	73.7±2.1	66.6±3.0	74.6±3.6	74.5±3.3	64.5±3.5
	FedDMVC (2023) [8]	81.1±0.9	81.9±0.9	73.7±1.4	69.9±1.5	72.5±2.9	58.9±2.6	63.1±0.4	62.6±0.5	48.4±0.3
	FCUIF (2024) [36]	85.3±0.3	83.2±0.3	75.7±0.5	72.8±1.2	70.3±1.5	64.4±1.8	67.2±0.8	64.4±0.9	53.5±0.8
	FMCSC (Ours)	**95.1±0.8**	**87.8±1.2**	**88.8±1.5**	**92.9±1.2**	**84.2±2.2**	**85.0±2.5**	**90.1±1.2**	**79.4±2.6**	**79.5±3.2**
BDGP	HCP-IMSC (2022) [24]	93.1±0.0	81.9±0.0	83.6±0.0	89.8±0.0	73.4±0.1	76.4±0.0	89.5±0.0	72.5±0.1	76.7±0.0
	IMVC-CBG (2022) [39]	37.9±0.4	21.0±1.0	10.4±0.1	37.2±0.1	21.0±0.0	7.8±0.0	36.9±0.1	20.4±0.1	6.4±0.0
	DSIMVC (2022) [37]	92.5±0.4	81.7±0.8	84.9±0.9	89.5±2.0	76.5±2.0	77.8±2.2	86.1±3.3	70.0±3.9	76.6±3.4
	LSIMVC (2022) [27]	44.1±0.5	23.7±0.4	5.7±0.2	39.2±1.3	19.7±0.5	4.8±0.2	35.3±1.6	14.9±1.3	2.8±0.4
	ProImp (2023) [22]	91.6±0.3	82.4±3.8	80.0±0.9	90.4±1.5	76.2±0.5	79.3±1.6	75.6±0.5	52.3±2.0	44.6±1.8
	JPLTD (2023) [29]	56.5±0.2	41.3±0.1	31.7±0.0	49.4±0.1	33.3±0.0	18.5±0.0	51.0±0.2	34.1±0.1	21.5±0.0
	CPSPAN (2023) [18]	78.7±0.6	58.3±1.3	58.6±1.6	57.3±1.3	50.3±2.3	39.4±3.7	52.4±1.5	34.7±1.4	27.1±2.1
	FedDMVC (2023) [8]	92.0±0.1	80.2±0.2	84.7±0.1	91.5±0.5	77.1±0.4	80.3±0.7	82.2±0.2	63.4±0.3	61.9±0.3
	FCUIF (2024) [36]	93.8±0.1	82.2±0.1	85.1±0.1	90.3±0.2	75.2±0.3	78.4±0.3	85.7±0.2	67.5±0.3	63.2±0.2
	FMCSC (Ours)	**94.5±0.8**	**83.9±1.2**	**86.8±1.5**	**91.9±1.2**	**77.3±2.2**	**81.0±2.5**	**90.0±1.2**	**73.3±2.6**	**76.8±3.2**
Multi-Fashion	HCP-IMSC (2022) [24]	70.6±0.1	67.4±0.1	57.7±0.1	67.1±0.1	64.7±0.1	53.1±0.1	59.9±0.7	56.4±0.9	41.8±1.1
	IMVC-CBG (2022) [39]	46.3±0.0	49.4±0.0	26.3±0.0	43.2±0.1	42.7±0.1	19.2±0.1	38.9±0.2	39.4±0.3	13.5±0.4
	DSIMVC (2022) [37]	82.7±1.3	83.6±1.1	74.5±1.1	77.7±1.4	76.7±0.8	66.8±0.8	76.7±1.7	75.8±1.5	66.4±1.4
	LSIMVC (2022) [27]	51.1±0.5	49.9±0.1	31.5±0.5	50.2±0.6	52.2±0.1	35.2±0.1	49.9±0.2	48.6±0.0	28.2±0.0
	ProImp (2023) [22]	69.1±0.4	66.3±0.3	55.2±0.8	69.0±0.1	64.6±0.2	52.5±0.2	53.9±2.4	50.7±1.5	27.8±1.6
	JPLTD (2023) [29]	44.6±0.0	43.4±0.0	36.9±0.1	37.2±0.1	36.6±0.1	29.5±0.1	25.8±0.1	25.1±0.1	16.5±0.1
	CPSPAN (2023) [18]	64.1±1.2	71.4±1.3	55.8±1.5	61.6±2.0	69.7±1.3	54.4±1.8	59.3±2.0	68.0±1.8	53.3±1.9
	FedDMVC (2023) [8]	67.7±0.3	74.6±0.8	58.0±0.6	66.6±0.4	65.3±0.7	54.3±0.7	57.6±0.7	58.5±0.8	43.2±0.7
	FCUIF (2024) [36]	70.7±0.5	79.4±0.5	63.1±0.4	68.4±0.6	71.5±0.4	59.2±0.5	62.5±0.6	61.3±0.5	45.6±0.5
	FMCSC (Ours)	**92.4±0.1**	**85.8±0.2**	**84.7±0.3**	**90.4±0.6**	**82.8±0.7**	**80.9±1.0**	**87.5±0.6**	**79.1±0.1**	**76.3±1.0**
NUSWIDE	HCP-IMSC (2022) [24]	36.1±0.0	8.5±0.0	6.7±0.0	35.3±0.1	8.2±0.0	6.3±0.0	31.3±0.0	6.0±0.1	4.7±0.1
	IMVC-CBG (2022) [39]	30.8±0.1	4.8±0.0	3.1±0.0	30.4±0.0	4.6±0.0	2.5±0.0	29.3±0.0	4.0±0.0	1.9±0.1
	DSIMVC (2022) [37]	51.1±1.3	25.3±0.8	23.4±0.8	50.6±0.8	22.2±0.7	20.4±0.6	46.7±0.5	18.3±0.6	16.0±0.4
	LSIMVC (2022) [27]	37.2±0.2	10.8±0.1	6.8±0.1	36.4±0.3	11.8±0.1	7.0±0.4	33.9±0.4	9.2±0.3	5.8±0.2
	ProImp (2023) [22]	38.4±0.1	11.1±0.0	8.3±0.1	37.1±0.4	10.5±0.1	7.6±0.2	34.3±0.7	8.0±0.0	6.1±0.0
	JPLTD (2023) [29]	53.0±0.2	25.9±0.2	23.7±0.1	51.5±0.5	22.5±1.0	21.5±0.8	50.0±0.2	19.4±0.1	14.1±0.1
	CPSPAN (2023) [18]	33.7±0.2	9.0±0.9	6.2±0.1	33.3±0.3	6.6±0.9	4.4±0.1	29.4±0.6	5.4±1.1	3.2±0.4
	FedDMVC (2023) [8]	41.7±0.2	14.4±0.1	12.3±0.1	37.5±0.7	9.8±0.9	7.8±0.5	32.6±0.2	5.8±0.2	4.3±0.2
	FCUIF (2024) [36]	45.2±0.3	15.0±0.3	14.1±0.2	40.2±0.5	10.0±0.6	9.2±0.5	38.2±0.4	9.6±0.3	8.2±0.3
	FMCSC (Ours)	**56.1±0.2**	**26.3±0.5**	**23.9±0.4**	**52.7±0.2**	**23.0±0.3**	**21.8±0.4**	**50.8±0.9**	**20.1±0.7**	**18.8±0.8**

Table 2: Ablation studies on four datasets when M/S=1:1.

	Components				MNIST-USPS			BDGP			Multi-Fashion			NUSWIDE		
	A	B	C	D	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Item-1		✓	✓	✓	91.12	82.24	81.63	69.84	48.71	46.29	87.80	79.86	76.61	42.16	12.37	11.16
Item-2	✓		✓	✓	60.82	39.74	34.28	61.04	30.76	30.14	56.04	34.22	27.75	42.08	10.15	9.06
Item-3	✓	✓		✓	88.66	76.85	76.84	87.40	68.96	71.09	82.58	74.29	68.82	45.86	15.06	12.64
Item-4	✓	✓	✓		89.52	78.12	78.31	85.44	64.15	67.26	88.14	79.61	76.91	47.54	15.20	13.79
Item-5	✓	✓	✓	✓	**92.93**	**84.18**	**85.02**	**91.92**	**77.29**	**80.95**	**90.36**	**82.81**	**80.89**	**52.74**	**22.97**	**21.81**

the number of communication rounds compared to FMCSC. Although the initial misalignment of the model on MNIST-USPS and Multi-Fashion can be mitigated through multiple communication rounds. Component A still plays a crucial role in our training process, facilitating consensus among clients during pre-training to alleviate model misalignment and accelerate convergence effectively. Item-3, which lacks component C, involves both single-view and multi-view clients carrying out the same operation of replacing their local models with global models. We observe a substantial improvement in Item-3 compared to Item-2, indicating that the success of model comparison in component C is attributed to high-quality global models. This achievement requires collaborative efforts from both multi-view and single-view clients. Additionally, the inclusion of the weighted aggregation in component D enhances the benefits of collaborative training among all clients.

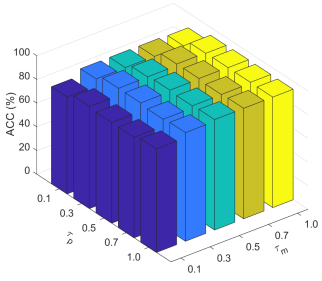


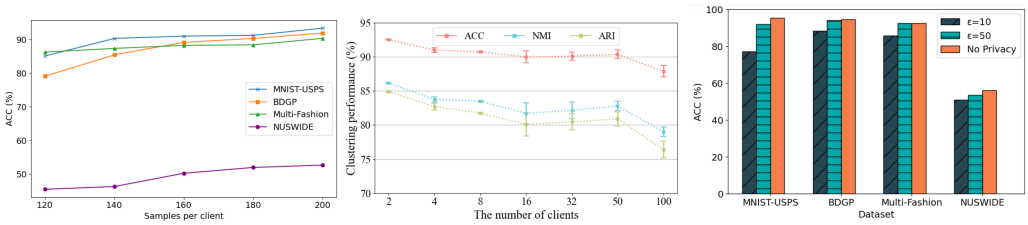
Figure 2: ACC vs. τ_m and τ_p.

Parameter Analysis. We investigate the sensitivity of our clustering performance on MNIST-USPS dataset to two primary hyperparameters in local-synergistic contrastive learning: τ_m and τ_p, as shown in Figure 2. The values of τ_m and τ_p are tuned within the range of 0.1, 0.3, 0.5, 0.7, 1. Our observations include: i) When τ_m and τ_p are set to small values, such as 0.1, the clustering performance of the proposed FMCSC decreases. This may be attributed to an excessive emphasis on view consistency, potentially resulting in an inseparable intrinsic feature space. ii) As the values of τ_m and τ_p increase, the clustering results gradually recover, and they exhibit insensitivity within the range of 0.3 to 1. Empirically, we set $\tau_m = \tau_p = 0.5$ for all datasets.

Qualitative Study on Model Misalignment. To further quantify the impact of model misalignment and the effectiveness of our proposed strategy, we visualize the model outputs, i.e., the consensus semantics $\mathbf{H}$, both without consensus pre-training and with consensus pre-training. Figure 3 displays the t-SNE [38] visualizations generated from a randomly selected multi-view client, where different colors represent different classes. In Figure 3 (a), we observe that the output of the global model without consensus pre-training shows mixed features that cannot be clearly distinguished. This confirms our viewpoint that direct parameter aggregation leads to model misalignment, manifested as feature confusion in the feature space. In contrast, Figure 3 (b) demonstrates that FMCSC produces more distinct and separable features, mitigating the negative impact of model aggregation.

(a) Without consensus. (b) With consensus.

Figure 3: Visualization on model misalignment.



(a) ACC vs. Samples per client. (b) ACC vs. Number of clients. (c) Privacy analysis.

Figure 4: (a) Effect of samples per client on generalization performance. (b) Scalability with the number of clients on Multi-Fashion. (c) Sensitivity under privacy constraints when $M/S = 2{:}1$.

4.3 Attributes of Federated Learning

Generalization Analysis. To check the validity of our theory for the proposed method, we investigate the impact of the number of samples per client and client participation rate on the generalization performance in the clustering task, as shown in Figure 4 (a) and Table 3. In this setting, the total number of clients is fixed. We find that: i) Increasing the number of samples per client effectively raises the total number of samples involved in training, which enhances the model's generalization performance. ii) As the proportion of participating clients increases, the accuracy of non-participating clients also improves, and the performance gap between participating and non-participating clients narrows. These observations are consistent with the theoretical understanding in Theorem 2, i.e., the generalization performance of the model is enhanced as both the number of samples per client and the number of participating clients increase.

Table 3: Effect of participation rates on generalization performance.

	Client types	Participating clients			Non-participating clients		
	Participation rate	50%	70%	90%	50%	70%	90%
Dataset	MNIST-USPS	90.64	91.71	92.97	89.11	91.67	92.06
	BDGP	88.32	89.48	92.73	85.92	86.63	87.65
	Multi-Fashion	87.58	89.46	89.81	87.31	87.48	88.73
	NUSWIDE	53.12	53.18	53.58	46.08	51.46	52.51

https://doi.org/10.52202/079017-1168

Number of Clients. We next consider the effects of changing the number of clients as shown in Figure 4 (b). It is observed that as the number of clients increases, the performance of FMCSC experiences a slight decline but remains generally stable. Only when the client number reaches 100, does a noticeable decline in performance occur, which is attributed to the insufficient number of samples within each client.

Privacy. FMCSC, by design, does not share any raw data between clients and the server. Only the model parameters on each client are shared with the server. To further protect client privacy, we adopt differential privacy [1] by adding noise to the model parameters uploaded from the client to the server. Figure 4 (c) illustrates the clustering accuracy of FMCSC under different privacy bounds ε. We observe that FMCSC achieves both high performance and privacy at $\varepsilon = 50$. However, as the level of noise increases at $\varepsilon = 10$, the performance of FMCSC unavoidably degrades.

5 Conclusion

In this paper, we propose FMCSC that can handle practical scenarios with heterogeneous hybrid views and explore the data cluster structures distributed on multiple clients. First, we propose cross-client consensus pre-training to align the local models on all clients to avoid their misalignment. Then, local-synergistic contrast and global-specific weighting aggregation are designed to bridge the client gap and the view gap across distributed clients and explore the cluster structures in multi-view data distributed on different clients. Theoretical analysis and extensive experiments demonstrate that FMCSC outperforms state-of-the-art methods across diverse heterogeneous hybrid views and various federated learning scenarios. In future work, we will use the method for more downstream tasks and some real-world situations, such as medical analysis and financial forecasting.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China (No. 62476052), Sichuan Science and Technology Program (No. 2024NSFSC1473), Central Guidance for Local Science and Technology Development Fund Projects (No. 2024ZYD0268), and Shenzhen Science and Technology Program (No. JCYJ20230807115959041).

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM CCS*, pages 308–318, 2016.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

[3] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NeurIPS*, pages 1–8, 2006.

[4] Xiao Cai, Hua Wang, Heng Huang, and Chris Ding. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12):i16–i24, 2012.

[5] Sicong Che, Zhaoming Kong, Hao Peng, Lichao Sun, Alex Leow, Yong Chen, and Lifang He. Federated multi-view learning for private medical data integration and analysis. *TIST*, 13(4):1–23, 2022.

[6] Jiayi Chen and Aidong Zhang. Fedmsplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *KDD*, pages 87–96, 2022.

[7] Man-Sheng Chen, Ling Huang, Chang-Dong Wang, and Dong Huang. Multi-view clustering in latent embedding space. In *AAAI*, pages 3513–3520, 2020.

[8] Xinyue Chen, Jie Xu, Yazhou Ren, Xiaorong Pu, Ce Zhu, Xiaofeng Zhu, Zhifeng Hao, and Lifang He. Federated deep multi-view clustering with global self-supervision. In *ACM MM*, pages 3498–3506, 2023.

[9] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM CIVR*, pages 1–9, 2009.

[10] Chenhang Cui, Yazhou Ren, Jingyu Pu, Jiawei Li, Xiaorong Pu, Tianyi Wu, Yutao Shi, and Lifang He. A novel approach for effective multi-view clustering with information-theoretic perspective. In *NeurIPS*, pages 1–13, 2023.

[11] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? In *NeurIPS*, pages 16937–16947, 2020.

[12] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *AISTATS*, pages 315–323, 2011.

[13] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *NeurIPS*, 6, 1993.

[14] Xingchen Hu, Jindong Qin, Yinghua Shen, Witold Pedrycz, Xinwang Liu, and Jiyuan Liu. An efficient federated multi-view fuzzy c-means clustering method. *TFS*, 32(4):1886–1899, 2023.

[15] Mingkai Huang, Hao Li, Bing Bai, Chang Wang, Kun Bai, and Fei Wang. A federated multi-view deep learning framework for privacy-preserving recommendations. *arXiv preprint arXiv:2008.10808*, 2020.

[16] Shudong Huang, Wei Shi, Zenglin Xu, Ivor W Tsang, and Jiancheng Lv. Efficient federated multi-view learning. *Pattern Recognition*, 131:108817, 2022.

[17] Weitian Huang, Sirui Yang, and Hongmin Cai. Generalized information-theoretic multi-view clustering. In *NeurIPS*, pages 1–13, 2023.

[18] Jiaqi Jin, Siwei Wang, Zhibin Dong, Xinwang Liu, and En Zhu. Deep incomplete multi-view clustering with cross-view partial sample and prototype alignment. In *CVPR*, pages 11600–11609, 2023.

[19] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, pages 5132–5143, 2020.

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[21] Rafał Latała and Krzysztof Oleszkiewicz. On the best constant in the khinchin-kahane inequality. *Studia Mathematica*, 109(1):101–104, 1994.

[22] Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. Incomplete multi-view clustering via prototype-based imputation. In *IJCAI*, pages 3911–3919, 2023.

[23] Shaojie Li and Yong Liu. Sharper generalization bounds for clustering. In *ICML*, pages 6392–6402, 2021.

[24] Zhenglai Li, Chang Tang, Xiao Zheng, Xinwang Liu, Wei Zhang, and En Zhu. High-order correlation preserved incomplete multi-view subspace clustering. *TIP*, 31:2067–2080, 2022.

[25] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. In *NeurIPS*, pages 1–28, 2023.

[26] Yi-Ming Lin, Yuan Gao, Mao-Guo Gong, Si-Jia Zhang, Yuan-Qiao Zhang, and Zhi-Yuan Li. Federated learning on multimodal data: A comprehensive survey. *MIR*, pages 1–15, 2023.

[27] Chengliang Liu, Zhihao Wu, Jie Wen, Yong Xu, and Chao Huang. Localized sparse incomplete multi-view clustering. *TMM*, 25:5539–5551, 2022.

 https://doi.org/10.52202/079017-1168

[28] Xinwang Liu, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu. Efficient and effective regularized incomplete multi-view clustering. *TPAMI*, 43(8):2634–2646, 2020.

[29] Wei Lv, Chao Zhang, Huaxiong Li, Xiuyi Jia, and Chunlin Chen. Joint projection learning and tensor decomposition-based incomplete multiview clustering. *TNNLS*, pages 1–12, 2023.

[30] J MacQueen. Classification and analysis of multivariate observations. In *BSMSP*, pages 281–297, 1967.

[31] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[34] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *ICML*, pages 5092–5101, 2019.

[35] Dong Qiao, Chris Ding, and Jicong Fan. Federated spectral clustering via secure similarity reconstruction. In *NeurIPS*, pages 1–36, 2023.

[36] Yazhou Ren, Xinyue Chen, Jie Xu, Jingyu Pu, Yonghao Huang, Xiaorong Pu, Ce Zhu, Xiaofeng Zhu, Zhifeng Hao, and Lifang He. A novel federated multi-view clustering method for unaligned and incomplete data fusion. *Information Fusion*, 108:1–10, 2024.

[37] Huayi Tang and Yong Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *ICML*, pages 21090–21110, 2022.

[38] Van, Laurens der Maaten, and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

[39] Siwei Wang, Xinwang Liu, Li Liu, Wenxuan Tu, Xinzhong Zhu, Jiyuan Liu, Sihang Zhou, and En Zhu. Highly-efficient incomplete large-scale multi-view clustering with consensus bipartite graph. In *CVPR*, pages 9776–9785, 2022.

[40] Song Wu, Yan Zheng, Yazhou Ren, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, and Lifang He. Self-weighted contrastive fusion for deep multi-view clustering. *TMM*, 2024.

[41] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[42] Jie Xu, Shuo Chen, Yazhou Ren, Xiaoshuang Shi, Heng Tao Shen, Gang Niu, and Xiaofeng Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. In *NeurIPS*, 2023.

[43] Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, and Xiaofeng Zhu. Investigating and mitigating the side effects of noisy views for self-supervised clustering algorithms in practical multi-view scenarios. In *CVPR*, pages 22957–22966, 2024.

[44] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *CVPR*, pages 16051–16060, 2022.

[45] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *CVPR*, pages 19863–19872, 2023.

[46] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *TPAMI*, 45(1):1055–1069, 2022.

[47] Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble. In *ICLR*, 2023.

[48] Fengda Zhang, Kun Kuang, Long Chen, Zhaoyang You, Tao Shen, Jun Xiao, Yin Zhang, Chao Wu, Fei Wu, Yueting Zhuang, et al. Federated unsupervised representation learning. *FITEE*, 24(8):1181–1193, 2023.

[49] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. *TPAMI*, 41(7):1774–1782, 2018.

[50] Handong Zhao, Hongfu Liu, and Yun Fu. Incomplete multi-modal visual data grouping. In *IJCAI*, pages 2392–2398, 2016.

[51] Huasong Zhong, Chong Chen, Zhongming Jin, and Xian-Sheng Hua. Deep robust clustering by contrastive learning. *arXiv preprint arXiv:2008.03030*, 2020.

https://doi.org/10.52202/079017-1168

We provide more details and results about our work in the appendices. Here are the contents:

- Appendix A: Framework of the proposed algorithm.
- Appendix B: More related work.
- Appendix C: Proofs of Theorem 1 and Theorem 2.
- Appendix D: More details about experimental settings.
- Appendix E: Additional experiment results.
- Appendix F: Broader impacts of our proposed method.
- Appendix G: Limitations of our proposed method.

A Framework of the Proposed Algorithm

Algorithm 1 outlines the execution flow for both clients and the server in FMCSC. Initially, cross-client consensus pre-training aligns the unsupervised local models of each client (Lines 1-2). During the training, feature contrastive learning is employed for multi-view clients to explore common semantics among different views (Lines 5-12). For single-view clients, model contrastive learning is designed between local models and global models, promoting the extraction of more generalized common semantics from client models (Lines 13-19). Subsequently, the server develops global-specific weighting aggregation to aggregate multiple high-quality models (Lines 21-23). Finally, complementary cluster structures are discovered using the global models across all clients (Line 25).

Algorithm 1 Federated Multi-view Clustering via Synergistic Contrast (FMCSC)

Input: Data with V views distributed among M multi-view clients and S single-view clients, communication rounds R, Local epoch E.
Output: Overall clustering results.

1: Pre-train all client models by Eqs. (1)-(2).
2: server receives pre-trained models from all clients and initializes the global models.
3: **while** not reaching R rounds **do**
4: **for** $c = 1$ to $(M + S)$ **do in parallel**
5: **if** $c \in [M]$ **then** ▷ Multi-view clients
6: **while** not reach the maximum iterations E **do**
7: Learn common semantics by Eqs. (4)-(5).
8: Optimize the total loss by Eq. (8).
9: **end while**
10: Optimize multiple global models by Eq. (6).
11: Upload $\{f_m^v(\cdot; \mathbf{w}_m^v)\}_{v=1}^V$ and $f_m(\cdot; \mathbf{w}_m)$ to the server.
12: **else if** $c \in [S]$ **then** ▷ Single-view clients
13: **while** not reach the maximum iterations E **do**
14: Learn common semantics by Eq. (7).
15: Optimize the total loss by Eq. (8).
16: **end while**
17: Upload $f_p\left(\cdot; \mathbf{w}_p^v\right)$ to the server.
18: **end if**
19: **end for** ▷ Server
20: Assigning weights to local models by Eq. (9).
21: Aggregate models among all clients by Eq. (10).
22: Distribute multiple global models to each client.
23: **end while**
24: Calculate the clustering results by Eq. (12).

B More Related Work

Multi-view clustering (MVC) methods leverage consistency and complementary information between multiple views to enhance clustering effectiveness. Based on their ability to handle missing data,

existing MVC methods can be classified into two categories. Complete multi-view clustering [7, 43, 49] which uncovers hidden patterns and structures by leveraging complete multi-view data for clustering. The success of existing complete multi-view clustering relies on the assumption of sample integrity across multiple views. However, in real-world scenarios, samples of multi-view are partially available due to data corruption or sensor failure, which leads to incomplete multi-view clustering studies [28, 46, 50]. Incomplete multi-view clustering via prototype-based imputation (ProImp) [22] employs a dual attention layer and a dual contrastive learning loss to learn view-specific prototypes and recover missing data. Cross-view partial sample and prototype alignment network (CPSPAN) [18] employs complete data to guide sample reconstruction and proposes shifted prototype alignment to calibrate prototype sets across views. However, the aforementioned MVC methods assume that multi-view data are stored within a single entity, thus lacking the concept of heterogeneous clients, and only addressing the hybrid views scenarios we mentioned.

C Theoretical Analysis

C.1 Proof of Theorem 1

In this part, we want to prove that minimizing contrastive loss $\mathcal{L}_c^m$ and $\mathcal{L}_c^p$ are equal to maximizing mutual information. The proof is motivated by [32, 51].

Proof. $\mathcal{L}_c^m$ and $\mathcal{L}_c^p$ as the contrastive loss, denoting the consistency objectives of the multi-view client m and single-view client p respectively, i.e.,

$$\mathcal{L}_c^m = -\frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}\sum_{i=1}^{|\mathcal{M}_m|}\log\frac{e^{sim(\mathbf{h}_i,\mathbf{h}_i^v)/\tau_m}}{\sum_{j\neq i}e^{sim\left(\mathbf{h}_i,\mathbf{h}_j^v\right)/\tau_m}},$$

$$\mathcal{L}_c^p = -\frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\log\frac{e^{sim\left(\mathbf{h}_i,\mathbf{h}_i^g\right)/\tau_p}}{e^{sim\left(\mathbf{h}_i,\mathbf{h}_i^g\right)/\tau_p}+e^{sim\left(\mathbf{h}_i,\mathbf{z}_i^v\right)/\tau_p}}.$$

When $j\neq i$, we assume that $p\left(\mathbf{h}_i,\mathbf{h}_j^v\right)=p\left(\mathbf{h}_i\right)p\left(\mathbf{h}_j^v\right)$, which means that $\frac{p\left(\mathbf{h}_i,\mathbf{h}_j^v\right)}{p(\mathbf{h}_i)p\left(\mathbf{h}_j^v\right)}=1$. Let $\mathcal{N}_i=\sum_{j=1}^{|\mathcal{M}_m|}\frac{p\left(\mathrm{h}_i,\mathrm{h}_j^v\right)}{p(\mathrm{h}_i)p\left(\mathrm{h}_j^v\right)}$, we can then

$$\begin{aligned}
I\left(\mathbf{H};\mathbf{H}^v\right) &= \sum_{i=1}^{|\mathcal{M}_m|}\sum_{j=1}^{|\mathcal{M}_m|}p\left(\mathbf{h}_i,\mathbf{h}_j^v\right)\log\frac{p\left(\mathbf{h}_i,\mathbf{h}_j^v\right)}{p\left(\mathbf{h}_i\right)p\left(\mathbf{h}_j^v\right)}\\
&= \sum_{i=1}^{|\mathcal{M}_m|}p\left(\mathbf{h}_i,\mathbf{h}_i^v\right)\log\frac{p\left(\mathbf{h}_i,\mathbf{h}_i^v\right)}{p\left(\mathbf{h}_i\right)p\left(\mathbf{h}_i^v\right)}+\sum_{i=1}^{|\mathcal{M}_m|}\sum_{j\neq i}p\left(\mathbf{h}_i,\mathbf{h}_j^v\right)\log\frac{p\left(\mathbf{h}_i,\mathbf{h}_j^v\right)}{p\left(\mathbf{h}_i\right)p\left(\mathbf{h}_j^v\right)}\\
&= \sum_{i=1}^{|\mathcal{M}_m|}p\left(\mathbf{h}_i,\mathbf{h}_i^v\right)\log\left(\frac{p\left(\mathbf{h}_i,\mathbf{h}_i^v\right)}{p\left(\mathbf{h}_i\right)p\left(\mathbf{h}_i^v\right)\cdot\mathcal{N}_i}\cdot\mathcal{N}_i\right)\\
&= \sum_{i=1}^{|\mathcal{M}_m|}p\left(\mathbf{h}_i,\mathbf{h}_i^v\right)\log\frac{\frac{p(\mathbf{h}_i,\mathbf{h}_i^v)}{p(\mathbf{h}_i)p\left(\mathbf{h}_i^v\right)}}{\mathcal{N}_i}+\sum_{i=1}^{|\mathcal{M}_m|}p\left(\mathbf{h}_i,\mathbf{h}_i^v\right)\log\mathcal{N}_i.
\end{aligned}$$

Since positive pairs are correlated, we have the estimate: $p\left(\mathbf{h}_i,\mathbf{h}_i^v\right)\geq p\left(\mathbf{h}_i\right)p\left(\mathbf{h}_i^v\right)$. In addition, we assume that there exists a constant $\delta_m\in(0,1)$ such that $p\left(\mathbf{h}_i^v\mid\mathbf{h}_i\right)>\delta_m$, $i=1,2,\cdots,|\mathcal{M}_m|$ holds. According to [42] and with the estimation, we have $p\left(\mathbf{h}_i\right)\approx\frac{1}{|\mathcal{M}_m|}$, $i=1,2,\cdots,|\mathcal{M}_m|$, and $e^{sim\left(\mathbf{h}_i,\mathbf{h}_j^v\right)/\tau_m}\propto\frac{p\left(\mathbf{h}_i,\mathbf{h}_j^v\right)}{p(\mathbf{h}_i)p\left(\mathbf{h}_j^v\right)}$, the following inequality holds:

https://doi.org/10.52202/079017-1168

$$
\begin{aligned}
\sum_{v=1}^{V} I\left(\mathbf{H}, \mathbf{H}^{v}\right) &= \sum_{v=1}^{V} \sum_{i=1}^{|\mathcal{M}_m|} p\left(\mathbf{h}_i, \mathbf{h}_i^{v}\right) \log \frac{\frac{p\left(\mathbf{h}_i, \mathbf{h}_i^{v}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{h}_i^{v}\right)}}{\mathcal{N}_i} + \sum_{v=1}^{V} \sum_{i=1}^{|\mathcal{M}_m|} p\left(\mathbf{h}_i, \mathbf{h}_i^{v}\right) \log \left(\sum_{j=1}^{|\mathcal{M}_m|} \frac{p\left(\mathbf{h}_i, \mathbf{h}_j^{v}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{h}_j^{v}\right)} \right) \\
&\approx \sum_{v=1}^{V} \sum_{i=1}^{|\mathcal{M}_m|} \frac{1}{|\mathcal{M}_m|} p\left(\mathbf{h}_i^{v} \mid \mathbf{h}_i\right) \log \frac{\frac{p\left(\mathbf{h}_i, \mathbf{h}_i^{v}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{h}_i^{v}\right)}}{\mathcal{N}_i} + \sum_{v=1}^{V} \log \left(|\mathcal{M}_m| - 1 + \frac{p\left(\mathbf{h}_i, \mathbf{h}_i^{v}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{h}_i^{v}\right)} \right) \\
&\geq \frac{\delta_m}{|\mathcal{M}_m|} \sum_{v=1}^{V} \sum_{i=1}^{|\mathcal{M}_m|} \log \frac{e^{sim\left(\mathbf{h}_i, \mathbf{h}_i^{v}\right)/\tau_m}}{\sum_{j \neq i} e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_j^{v}\right)/\tau_m} + e^{sim\left(\mathbf{h}_i, \mathbf{h}_i^{v}\right)/\tau_m}} + V \log |\mathcal{M}_m| \\
&\geq V \log |\mathcal{M}_m| - \delta_m \mathcal{L}_c^{m}.
\end{aligned}
$$

Similarly, we assume that there exists a constant $\delta_p \in (0,1)$ such that $p\left(\mathbf{h}_i^{g} \mid \mathbf{h}_i\right) = p\left(\mathbf{z}_i^{v} \mid \mathbf{h}_i\right) > \delta_p$, $i = 1, 2, \cdots, |\mathcal{S}_p|$ holds. And we have $p\left(\mathbf{h}_i\right) \approx \frac{1}{|\mathcal{S}_p|}$, $e^{sim\left(\mathbf{h}_i, \mathbf{h}_i^{g}\right)/\tau_p} \propto \frac{p\left(\mathbf{h}_i, \mathbf{h}_i^{g}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{h}_i^{g}\right)}$ and $e^{sim\left(\mathbf{h}_i, \mathbf{z}_i^{v}\right)/\tau_p} \propto \frac{p\left(\mathbf{h}_i, \mathbf{z}_i^{v}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{z}_i^{v}\right)}$, the following inequality holds:

$$
\begin{aligned}
I\left(\mathbf{H}, \mathbf{H}^{g}\right) - I\left(\mathbf{H}, \mathbf{Z}^{v}\right) =& \sum_{i=1}^{|\mathcal{S}_p|} p\left(\mathbf{h}_i, \mathbf{h}_i^{g}\right) \log \frac{p\left(\mathbf{h}_i, \mathbf{h}_i^{g}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{h}_i^{g}\right)} - \sum_{i=1}^{|\mathcal{S}_p|} p\left(\mathbf{h}_i, \mathbf{z}_i^{v}\right) \log \frac{p\left(\mathbf{h}_i, \mathbf{z}_i^{v}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{z}_i^{v}\right)} \\
\approx& \sum_{i=1}^{|\mathcal{S}_p|} \frac{2}{|\mathcal{S}_p|} p\left(\mathbf{h}_i^{g} \mid \mathbf{h}_i\right) \log \frac{p\left(\mathbf{h}_i, \mathbf{h}_i^{g}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{h}_i^{g}\right)} - \sum_{i=1}^{|\mathcal{S}_p|} \frac{1}{|\mathcal{S}_p|} p\left(\mathbf{h}_i^{g} \mid \mathbf{h}_i\right) \log \frac{p\left(\mathbf{h}_i, \mathbf{h}_i^{g}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{h}_i^{g}\right)} \\
& - \sum_{i=1}^{|\mathcal{S}_p|} \frac{1}{|\mathcal{S}_p|} p\left(\mathbf{z}_i^{v} \mid \mathbf{h}_i\right) \log \frac{p\left(\mathbf{h}_i, \mathbf{z}_i^{v}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{z}_i^{v}\right)} \\
\leq& \frac{2\delta_p}{|\mathcal{S}_p|} \sum_{i=1}^{|\mathcal{S}_p|} \log \frac{p\left(\mathbf{h}_i, \mathbf{h}_i^{g}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{h}_i^{g}\right)} - \frac{2\delta_p}{|\mathcal{S}_p|} \sum_{i=1}^{|\mathcal{S}_p|} \log \left(\frac{p\left(\mathbf{h}_i, \mathbf{h}_i^{g}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{h}_i^{g}\right)} + \frac{p\left(\mathbf{h}_i, \mathbf{z}_i^{v}\right)}{p\left(\mathbf{h}_i\right) p\left(\mathbf{z}_i^{v}\right)} \right) \\
\leq& \frac{2\delta_p}{|\mathcal{S}_p|} \sum_{i=1}^{|\mathcal{S}_p|} \log \frac{e^{sim\left(\mathbf{h}_i, \mathbf{h}_i^{g}\right)/\tau_p}}{e^{sim\left(\mathbf{h}_i, \mathbf{h}_i^{g}\right)/\tau_p} + e^{sim\left(\mathbf{h}_i, \mathbf{z}_i^{v}\right)/\tau_p}} \\
\leq& -2\delta_p \mathcal{L}_c^{p}.
\end{aligned}
$$

□

C.2 Proof of Theorem 2

We consider a heterogeneous federated learning setting where M multi-view clients and S single-view clients. For each multi-view client, we define the empirical risk and its expectation as $\widehat{\mathcal{L}}_m(f)$ and $\mathcal{L}_m(f)$, respectively. Similarly, for each single-view client, the empirical risk and its expectation are denoted as $\widehat{\mathcal{L}}_p(f)$ and $\mathcal{L}_p(f)$. Let $f : \mathcal{X} \to \mathbb{R}^{D_v} \times \mathbb{R}^{d_v} \times \mathbb{R}^{d_v}$ denotes the function that maps input samples into reconstruction samples, low-level features and high-level features. Then, the reconstruction samples, low-level features and high-level features are given by $\hat{\mathbf{x}}_i^{v} := f_x\left(\mathbf{x}_i^{v}\right) \in \mathbb{R}^{D_v}$, $\mathbf{z}_i^{v} := f_z^{v}\left(\mathbf{x}_i^{v}\right) \in \mathbb{R}^{d_v}$, and $\mathbf{h}_i^{v} := f_h^{v}\left(\mathbf{x}_i^{v}\right) \in \mathbb{R}^{d_v}$, respectively. Additionally, we define $\left\{\mathbf{x}_i^{v}\right\}_{v=1}^{V} := \mathbf{x}_i$, then common semantics are denoted as $\mathbf{h}_i := f_h^{0}\left(\mathbf{x}_i\right) \in \mathbb{R}^{d}$. To prove Theorem 2, we first introduce the following three lemmas.

Lemma 1. *We define the empirical risk $\widehat{\mathcal{L}}_m(f)$ for multi-view client m as*

$$
\begin{aligned}
\widehat{\mathcal{L}}_m(f) &= \frac{1}{|\mathcal{M}_m|} \sum_{v=1}^{V} \sum_{i=1}^{|\mathcal{M}_m|} \left[\left\| \mathbf{x}_i^{v} - f_x\left(\mathbf{x}_i^{v}\right) \right\|^2 - \log \frac{e^{sim\left(\mathbf{h}_i, \mathbf{h}_i^{v}\right)/\tau_m}}{\sum_{j \neq i} e^{sim\left(\mathbf{h}_i, \mathbf{h}_j^{v}\right)/\tau_m}} \right] \\
&= \frac{1}{|\mathcal{M}_m|} \sum_{v=1}^{V} \sum_{i=1}^{|\mathcal{M}_m|} \left[\left\| \mathbf{x}_i^{v} - f_x\left(\mathbf{x}_i^{v}\right) \right\|^2 - \log \frac{e^{sim\left(f_h^{0}\left(\mathbf{x}_i\right), f_h^{v}\left(\mathbf{x}_i^{v}\right)\right)/\tau_m}}{\sum_{j \neq i} e^{sim\left(f_h^{0}\left(\mathbf{x}_i\right), f_h^{v}\left(\mathbf{x}_j^{v}\right)\right)/\tau_m}} \right].
\end{aligned}
$$

Let $\mathcal{L}_m(f)$ be the expectation of $\widehat{\mathcal{L}}_m(f)$. Suppose that for any $\mathbf{x} \in \mathcal{X}$ and $f \in \mathcal{F}$, there exists $D < \infty$ such that $\|\mathbf{x}\|, \|f_x(\mathbf{x})\|, \|f_h^v(\mathbf{x})\|, \left\|f_h^0(\mathbf{x})\right\| \in [0, D]$ hold. With probability $1-\delta$ for any $f \in \mathcal{F}$, the following inequality holds

$$\mathcal{L}_m(f) \leq \widehat{\mathcal{L}}_m(f) + \frac{12VD^2}{\sqrt{|\mathcal{M}_m|}} + 9VD^2\sqrt{\frac{\log\frac{1}{\delta}}{2|\mathcal{M}_m|}}.$$

Proof. This proof is inspired by [37]. For simplicity, we define $f_h(\mathbf{x}_i, \mathbf{x}_j) := sim\left(f_h^0(\mathbf{x}_i), f_h^v(\mathbf{x}_j^v)\right)/\tau_m$ and $f_h(\mathbf{x}_i) := sim\left(f_h^0(\mathbf{x}_i), f_h^v(\mathbf{x}_i^v)\right)/\tau_m$, respectively. Then, the empirical risk and its expectation can be formulated as

$$\begin{aligned}
\widehat{\mathcal{L}}_m(f) &= \frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}\sum_{i=1}^{|\mathcal{M}_m|}\left[\|\mathbf{x}_i^v - f_x(\mathbf{x}_i^v)\|^2 - \log\frac{e^{f_h(\mathbf{x}_i)}}{\sum_{j\neq i}e^{f_h(\mathbf{x}_i,\mathbf{x}_j)}}\right] \\
&= \frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}\sum_{i=1}^{|\mathcal{M}_m|}\left[\|\mathbf{x}_i^v - f_x(\mathbf{x}_i^v)\|^2 + \log\left(\sum_{j\neq i}\exp\{f_h(\mathbf{x}_i,\mathbf{x}_j)\}\right) - f_h(\mathbf{x}_i)\right] \\
&= \frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}\sum_{i=1}^{|\mathcal{M}_m|}\left[\|\mathbf{x}_i^v - f_x(\mathbf{x}_i^v)\|^2 - f_h(\mathbf{x}_i)\right] + \frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}\sum_{i=1}^{|\mathcal{M}_m|}\log\sum_{j\neq i}\exp\{f_h(\mathbf{x}_i,\mathbf{x}_j)\},
\end{aligned}$$

and

$$\mathcal{L}_m(f) = \sum_{v=1}^{V}\mathbb{E}_{\mathbf{x}}\left[\|x_i^v - f_x(x_i^v)\|^2 - f_h(\mathbf{x}_i)\right] + \sum_{v=1}^{V}\mathbb{E}_{\mathbf{x}}\left[\log\sum_{j\neq i}\exp\{f_h(\mathbf{x}_i,\mathbf{x}_j)\}\right].$$

Let $\bar{\mathcal{M}}_m$ be the sample set that different from $\mathcal{M}_m$ by only one set of data $\bar{\mathbf{x}}_r := \{\bar{\mathbf{x}}_r^v\}_{v=1}^V$. The empirical risk of the function f on $\bar{\mathcal{M}}_m$ is denoted as $\widehat{\mathcal{L}}'_m(f)$. We have

$$\begin{aligned}
&\left|\sup_{f\in\mathcal{F}}|\mathcal{L}_m(f) - \widehat{\mathcal{L}}_m(f)| - \sup_{f\in\mathcal{F}}|\mathcal{L}_m(f) - \widehat{\mathcal{L}}'_m(f)|\right| \\
\leq &\sup_{f\in\mathcal{F}}\left|\widehat{\mathcal{L}}_m(f) - \widehat{\mathcal{L}}'_m(f)\right| \\
\leq &\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}\left(\|\mathbf{x}_r^v - f_x(\mathbf{x}_r^v)\|^2 - \|\bar{\mathbf{x}}_r^v - f_x(\bar{\mathbf{x}}_r^v)\|^2 - (f_h(\mathbf{x}_r) - f_h(\bar{\mathbf{x}}_r))\right)\right| \\
&+ \sup_{f\in\mathcal{F}}\left|\frac{2}{|\mathcal{M}_m|}\sum_{v=1}^{V}\left(\log\sum_{j\neq i}\exp\{f_h(\mathbf{x}_r,\mathbf{x}_j)\} - \log\sum_{j\neq i}\exp\{f_h(\bar{\mathbf{x}}_r,\mathbf{x}_j)\}\right)\right| \\
\leq &\sup_{f\in\mathcal{F}}\frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}\left|\|\mathbf{x}_r^v\|^2 - \|\bar{\mathbf{x}}_r^v\|^2 + \|f_x(\mathbf{x}_r^v)\|^2 - \|f_x(\bar{\mathbf{x}}_r^v)\|^2 + 2\|\mathbf{x}_r^v\|\|f_x(\mathbf{x}_r^v)\| + 2\|\bar{\mathbf{x}}_r^v\|\|f_x(\bar{\mathbf{x}}_r^v)\|\right| \\
&+ \sup_{f\in\mathcal{F}}\frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}|f_h(\mathbf{x}_r) - f_h(\bar{\mathbf{x}}_r)| + \sup_{f\in\mathcal{F}}\frac{2}{|\mathcal{M}_m|(|\mathcal{M}_m|-1)}\sum_{v=1}^{V}\left|\sum_{j\neq i}(f_h(\mathbf{x}_r,\mathbf{x}_j) - f_h(\bar{\mathbf{x}}_r,\mathbf{x}_j))\right| \\
\leq &\frac{9VD^2}{|\mathcal{M}_m|}.
\end{aligned}$$

Then, we analyze the upper bound of the expectation term $\mathbb{E}\sup_{f\in\mathcal{F}}|\widehat{\mathcal{L}}_m(f) - \mathcal{L}_m(f)|$. Let $\sigma_1, \ldots, \sigma_{|\mathcal{M}_m|}$ be i.i.d. independent random variables taking values in $\{-1, 1\}$ and $\bar{\mathcal{M}}_m$ be the independent copy of $\mathcal{M}_m$. We have

https://doi.org/10.52202/079017-1168

$$
\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}}|\widehat{\mathcal{L}}_m(f)-\mathcal{L}_m(f)| \\
\leq&\mathbb{E}_{\mathcal{M}_m,\bar{\mathcal{M}}_m}\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}\sum_{i=1}^{|\mathcal{M}_m|}\left(\left\|\mathbf{x}_i^v-f_x\left(\mathbf{x}_i^v\right)\right\|^2-\left\|\bar{\mathbf{x}}_i^v-f_x\left(\bar{\mathbf{x}}_i^v\right)\right\|^2\right)\right| \\
&+\mathbb{E}_{\mathcal{M}_m,\bar{\mathcal{M}}_m}\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}\sum_{i=1}^{|\mathcal{M}_m|}\left(f_h\left(\mathbf{x}_i\right)-f_h\left(\bar{\mathbf{x}}_i\right)\right)\right| \\
&+\mathbb{E}_{\mathcal{M}_m,\bar{\mathcal{M}}_m}\sup_{f\in\mathcal{F}}\left|\frac{2}{|\mathcal{M}_m|}\sum_{v=1}^{V}\sum_{i=1}^{|\mathcal{M}_m|}\left(\log\sum_{j\neq i}\exp\left\{f_h\left(\mathbf{x}_i,\mathbf{x}_j\right)\right\}-\log\sum_{j\neq i}\exp\left\{f_h\left(\bar{\mathbf{x}}_i,\mathbf{x}_j\right)\right\}\right)\right| \\
\leq&2\mathbb{E}_{\mathcal{M}_m,\sigma}\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}\sum_{i=1}^{|\mathcal{M}_m|}\sigma_i\left\|\mathbf{x}_i^v-f_x\left(\mathbf{x}_i^v\right)\right\|^2\right|+2\mathbb{E}_{\mathcal{M}_m,\sigma}\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{M}_m|}\sum_{v=1}^{V}\sum_{i=1}^{|\mathcal{M}_m|}\sigma_i f_h\left(\mathbf{x}_i\right)\right| \\
&+2\mathbb{E}_{\mathcal{M}_m,\sigma}\sup_{f\in\mathcal{F}}\left|\frac{2}{|\mathcal{M}_m|\left(|\mathcal{M}_m|-1\right)}\sum_{v=1}^{V}\sum_{i=1}^{|\mathcal{M}_m|}\sum_{j\neq i}\sigma_i f_h\left(\mathbf{x}_i,\mathbf{x}_j\right)\right| \\
\leq&2V\max_v\mathbb{E}_{\mathcal{M}_m,\sigma}\sup_{f\in\mathcal{F}}\left(\frac{1}{|\mathcal{M}_m|}\sum_{i=1}^{|\mathcal{M}_m|}\left[\left\|\mathbf{x}_i^v-f_x\left(\mathbf{x}_i^v\right)\right\|^2\right]^2\right)^{\frac{1}{2}}+2V\max_v\mathbb{E}_{\mathcal{M}_m,\sigma}\sup_{f\in\mathcal{F}}\left(\frac{1}{|\mathcal{M}_m|}\sum_{i=1}^{|\mathcal{M}_m|}\left[f_h\left(\mathbf{x}_i\right)\right]^2\right)^{\frac{1}{2}} \\
&+4V\max_v\mathbb{E}_{\mathcal{M}_m,\sigma}\sup_{f\in\mathcal{F}}\left(\frac{1}{|\mathcal{M}_m|\left(|\mathcal{M}_m|-1\right)}\sum_{i=1}^{|\mathcal{M}_m|}\sum_{j\neq i}\left[f_h\left(\mathbf{x}_i\right)\right]^2\right)^{\frac{1}{2}}, \\
\leq&\frac{12VD^2}{\sqrt{|\mathcal{M}_m|}},
\end{aligned}
$$

where the second last inequality is obtained by the Khintchine-Kahane inequality[21]. Thus, according to the McDiarmid inequality [31], with probability at least $1-\delta$ for any $f\in\mathcal{F}$, we have

$$
\mathcal{L}_m(f)\leq\widehat{\mathcal{L}}_m(f)+\frac{12VD^2}{\sqrt{|\mathcal{M}_m|}}+9VD^2\sqrt{\frac{\log\frac{1}{\delta}}{2\left|\mathcal{M}_m\right|}}.
$$

□

Due to the presence of view gaps among different views, let the data from the v-th view follow the distribution $\mathcal{D}_v$, and the multi-view data follow the distribution $\mathcal{D}$. We analyze the generalization bounds for $f_m^v\left(\cdot;\mathbf{w}_m^v\right)$ in multi-view client m, which is built upon prior works from domain adaptation.

Lemma 2 (Generalization Bounds for Domain Adaptation [2, 3])**.** *Consider the v-th view data domain $\mathcal{D}_v$ and the multi-view data domain $\mathcal{D}$, respectively. Given a feature extraction function $\mathcal{R}:\mathcal{X}\mapsto\mathcal{H}$ that shared between $\mathcal{D}_v$ and $\mathcal{D}$. Let $\mathcal{F}$ be a set of hypothesis with VC-dimension d. Then, for every $f\in\mathcal{F}$, with probability at least $1-\delta$:*

$$
\mathcal{L}_m(f)\leq\mathcal{L}_{m^v}(f)+\sqrt{\frac{4}{|\mathcal{M}_m|}\left(d\log\frac{2e\left|\mathcal{M}_m\right|}{d}+\log\frac{4}{\delta}\right)}+d_{\mathcal{F}}\left(\tilde{\mathcal{D}}_v,\tilde{\mathcal{D}}\right)+\lambda_v,
$$

where $d_{\mathcal{F}}\left(\tilde{\mathcal{D}}_v,\tilde{\mathcal{D}}\right)$ denotes the divergence measured over a symmetric-difference hypothesis space. $\tilde{\mathcal{D}}_v$ and $\tilde{\mathcal{D}}$ are the induced distributions of $\mathcal{D}_v$ and $\mathcal{D}$ under $\mathcal{R}$, respectively, s.t. $\mathbb{E}_{h\sim\tilde{\mathcal{D}}_v}[\mathcal{B}(h)]=\mathbb{E}_{x\sim\mathcal{D}_v}[\mathcal{B}(\mathcal{R}(x))]$ given a probability event $\mathcal{B}$, and so for $\tilde{\mathcal{D}}$. $\lambda_v:=\min_f\mathcal{L}_{m^v}(f)+\mathcal{L}_m(f)$ denotes an oracle performance.

Lemma 3. *We define the empirical risk $\widehat{\mathcal{L}}_p(f)$ for single-view client p as*

$$\begin{aligned}\widehat{\mathcal{L}}_p(f) &= \frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\left[\left\|\mathbf{x}_i^v - f_x\left(\mathbf{x}_i^v\right)\right\|^2 - \log\frac{e^{sim\left(\mathbf{h}_i,\mathbf{h}_i^g\right)/\tau_p}}{e^{sim\left(\mathbf{h}_i,\mathbf{h}_i^g\right)/\tau_p}+e^{sim\left(\mathbf{h}_i,\mathbf{z}_i^v\right)/\tau_p}}\right] \\ &= \frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\left[\left\|\mathbf{x}_i^v - f_x\left(\mathbf{x}_i^v\right)\right\|^2 - \log\frac{e^{sim\left(f_h^0(\mathbf{x}_i),f_g^0(\mathbf{x}_i)\right)/\tau_p}}{e^{sim\left(f_h^0(\mathbf{x}_i),f_g^0(\mathbf{x}_i)\right)/\tau_p}+e^{sim\left(f_h^0(\mathbf{x}_i),f_z^v\left(\mathbf{x}_i^v\right)\right)/\tau_p}}\right].\end{aligned}$$

Let $\mathcal{L}_p(f)$ be the expectation of $\widehat{\mathcal{L}}_p(f)$. Suppose that for any $\mathbf{x}\in\mathcal{X}$ and $f\in\mathcal{F}$, there exists $D<\infty$ such that $\|\mathbf{x}\|, \|f_x(\mathbf{x})\|, \|f_z^v(\mathbf{x})\|, \left\|f_h^0(\mathbf{x})\right\|, \left\|f_g^0(\mathbf{x})\right\| \in [0,D]$ hold. With probability $1-\delta$ for any $f\in\mathcal{F}$, the following inequality holds

$$\mathcal{L}_p(f) \le \widehat{\mathcal{L}}_p(f) + \frac{10D^2}{\sqrt{|\mathcal{S}_p|}} + 8D^2\sqrt{\frac{\log\frac{1}{\delta}}{2\,|\mathcal{S}_p|}}.$$

Proof. For simplicity, we define $f_g(\mathbf{x}_i) := sim\left(f_h^0(\mathbf{x}_i), f_g^0(\mathbf{x}_i)\right)/\tau_p$ and $f_{h,z}(\mathbf{x}_i) := sim\left(f_h^0(\mathbf{x}_i), f_z^v(\mathbf{x}_i^v)\right)/\tau_p$, respectively. Then, the empirical risk and its expectation can be formulated as

$$\begin{aligned}\widehat{\mathcal{L}}_p(f) &= \frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\left[\left\|\mathbf{x}_i^v - f_x\left(\mathbf{x}_i^v\right)\right\|^2 - \log\frac{e^{f_g(\mathbf{x}_i)}}{e^{f_g(\mathbf{x}_i)}+e^{f_{h,z}(\mathbf{x}_i)}}\right] \\ &= \frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\left[\left\|\mathbf{x}_i^v - f_x\left(\mathbf{x}_i^v\right)\right\|^2 + \log\left(e^{f_g(\mathbf{x}_i)}+e^{f_{h,z}(\mathbf{x}_i)}\right) - f_g\left(\mathbf{x}_i\right)\right] \\ &= \frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\left[\left\|\mathbf{x}_i^v - f_x\left(\mathbf{x}_i^v\right)\right\|^2 - f_g\left(\mathbf{x}_i\right)\right] + \frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\log\left(e^{f_g(\mathbf{x}_i)}+e^{f_{h,z}(\mathbf{x}_i)}\right),\end{aligned}$$

and

$$\mathcal{L}_p(f) = \mathbb{E}_{\mathbf{x}}\left[\left\|x_i^v - f_x\left(x_i^v\right)\right\|^2 - f_h\left(\mathbf{x}_i\right)\right] + \mathbb{E}_{\mathbf{x}}\left[\log\left(e^{f_g(\mathbf{x}_i)}+e^{f_{h,z}(\mathbf{x}_i)}\right)\right].$$

Let $\bar{\mathcal{S}}_p$ be the sample set that different from $\mathcal{S}_p$ by only one sample point $\bar{\mathbf{x}}_r$. The empirical risk of the function f on $\bar{\mathcal{S}}_p$ is denoted as $\widehat{\mathcal{L}}'_p(f)$. We have

$$\begin{aligned}&\left|\sup_{f\in\mathcal{F}}\left|\mathcal{L}_p(f)-\widehat{\mathcal{L}}_p(f)\right| - \sup_{f\in\mathcal{F}}\left|\mathcal{L}_p(f)-\widehat{\mathcal{L}}'_p(f)\right|\right| \\ \le& \sup_{f\in\mathcal{F}}\left|\widehat{\mathcal{L}}_p(f)-\widehat{\mathcal{L}}'_p(f)\right| \\ \le& \sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{S}_p|}\left(\left\|\mathbf{x}_r^v - f_x\left(\mathbf{x}_r^v\right)\right\|^2 - \left\|\bar{\mathbf{x}}_r^v - f_x\left(\bar{\mathbf{x}}_r^v\right)\right\|^2 - \left(f_g\left(\mathbf{x}_r\right) - f_g\left(\bar{\mathbf{x}}_r\right)\right)\right)\right| \\ &+ \sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{S}_p|}\left(\log\left(e^{f_g(\mathbf{x}_r)}+e^{f_{h,z}(\mathbf{x}_r)}\right) - \log\left(e^{f_g(\bar{\mathbf{x}}_r)}+e^{f_{h,z}(\bar{\mathbf{x}}_r)}\right)\right)\right| \\ \le& \sup_{f\in\mathcal{F}}\frac{1}{|\mathcal{S}_p|}\left|\left\|\mathbf{x}_r^v\right\|^2 - \left\|\bar{\mathbf{x}}_r^v\right\|^2 + \left\|f_x\left(\mathbf{x}_r^v\right)\right\|^2 - \left\|f_x\left(\bar{\mathbf{x}}_r^v\right)\right\|^2 + 2\left\|\mathbf{x}_r^v\right\|\left\|f_x\left(\mathbf{x}_r^v\right)\right\| + 2\left\|\bar{\mathbf{x}}_r^v\right\|\left\|f_x\left(\bar{\mathbf{x}}_r^v\right)\right\|\right| \\ &+ \sup_{f\in\mathcal{F}}\frac{1}{|\mathcal{S}_p|}\left|f_g\left(\mathbf{x}_r\right) - f_g\left(\bar{\mathbf{x}}_r\right)\right| + \sup_{f\in\mathcal{F}}\frac{1}{2\,|\mathcal{S}_p|}\left|\left(f_g\left(\mathbf{x}_r\right)+f_{h,z}\left(\mathbf{x}_r\right)\right) - \left(f_g\left(\bar{\mathbf{x}}_r\right)+f_{h,z}\left(\bar{\mathbf{x}}_r\right)\right)\right| \\ \le& \frac{8D^2}{|\mathcal{S}_p|}.\end{aligned}$$

https://doi.org/10.52202/079017-1168

Then, we analyze the upper bound of the expectation term $\mathbb{E}\sup_{f\in\mathcal{F}}|\widehat{\mathcal{L}}_p(f)-\mathcal{L}_p(f)|$. Let $\sigma_1,\ldots,\sigma_{|\mathcal{S}_p|}$ be i.i.d. independent random variables taking values in $\{-1,1\}$ and $\bar{\mathcal{S}}_p$ be the independent copy of $\mathcal{S}_p$. We have

$$
\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}}|\widehat{\mathcal{L}}_p(f)-\mathcal{L}_p(f)| \\
\leq&\mathbb{E}_{\mathcal{S}_p,\bar{\mathcal{S}}_p}\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\left(\left\|\mathbf{x}_i^v-f_x\left(\mathbf{x}_i^v\right)\right\|^2-\left\|\bar{\mathbf{x}}_i^v-f_x\left(\bar{\mathbf{x}}_i^v\right)\right\|^2\right)\right|+\mathbb{E}_{\mathcal{S}_p,\bar{\mathcal{S}}_p}\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\left(f_g\left(\mathbf{x}_i\right)-f_g\left(\bar{\mathbf{x}}_i\right)\right)\right| \\
&+\mathbb{E}_{\mathcal{S}_p,\bar{\mathcal{S}}_p}\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\left(\log\left(e^{f_g(\mathbf{x}_i)}+e^{f_{h,z}(\mathbf{x}_i)}\right)-\log\left(e^{f_g(\bar{\mathbf{x}}_i)}+e^{f_{h,z}(\bar{\mathbf{x}}_i)}\right)\right)\right| \\
\leq&2\mathbb{E}_{\mathcal{S}_p,\sigma}\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\sigma_i\left\|\mathbf{x}_i^v-f_x\left(\mathbf{x}_i^v\right)\right\|^2\right|+2\mathbb{E}_{\mathcal{S}_p,\sigma}\sup_{f\in\mathcal{F}}\left|\frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\sigma_i f_g\left(\mathbf{x}_i\right)\right| \\
&+2\mathbb{E}_{\mathcal{S}_p,\sigma}\sup_{f\in\mathcal{F}}\left|\frac{1}{2\,|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\sigma_i\left(f_g\left(\mathbf{x}_i\right)+f_{h,z}\left(\mathbf{x}_i\right)\right)\right| \\
\leq&2\mathbb{E}_{\mathcal{S}_p,\sigma}\sup_{f\in\mathcal{F}}\left(\frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\left[\left\|\mathbf{x}_i^v-f_x\left(\mathbf{x}_i^v\right)\right\|^2\right]^2\right)^{\frac{1}{2}}+2\mathbb{E}_{\mathcal{S}_p,\sigma}\sup_{f\in\mathcal{F}}\left(\frac{1}{|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\left[f_g\left(\mathbf{x}_i\right)\right]^2\right)^{\frac{1}{2}} \\
&+2\mathbb{E}_{\mathcal{S}_p,\sigma}\sup_{f\in\mathcal{F}}\left(\frac{1}{2\,|\mathcal{S}_p|}\sum_{i=1}^{|\mathcal{S}_p|}\left(\left[f_g\left(\mathbf{x}_i\right)\right]^2+\left[f_{h,z}\left(\mathbf{x}_i\right)\right]^2\right)\right)^{\frac{1}{2}}, \\
\leq&\frac{10D^2}{\sqrt{|\mathcal{S}_p|}}.
\end{aligned}
$$

Thus, according to McDiarmid inequality [31], with probability at least $1-\delta$ for any $f\in\mathcal{F}$, we have

$$
\mathcal{L}_p(f)\leq\widehat{\mathcal{L}}_p(f)+\frac{10D^2}{\sqrt{|\mathcal{S}_p|}}+8D^2\sqrt{\frac{\log\frac{1}{\delta}}{2\,|\mathcal{S}_p|}}.
$$

□

Our objective is to collaboratively train all clients to obtain multiple heterogeneous global models capable of handling different view types. This is manifested in the optimization of multiple global objective functions. For the global objective of handling multiple view types simultaneously, expected risk is defined as $\mathcal{L}_M(f)$, typically optimized in the form of empirical risk minimization, defined as:

$$
\widehat{\mathcal{L}}_M(f)=\frac{1}{M}\sum_{m=1}^{M}\widehat{\mathcal{L}}_m(f).
$$

Similarly, for handling individual view types, such as the v-th view type, the global objective entails defining the expected risk as $\mathcal{L}_{S^v}(f)$, with empirical risk defined as:

$$
\widehat{\mathcal{L}}_{S^v}(f)=\frac{1}{S^v}\sum_{p\in[S^v]}\widehat{\mathcal{L}}_p(f)+\frac{1}{M}\sum_{m=1}^{M}\widehat{\mathcal{L}}_{m^v}(f).
$$

Now we give the proof of Theorem 2.

Proof. According to Lemma 1, with probability at least $1-\delta$ for any $f \in \mathcal{F}$, we have

$$\mathcal{L}_M(f) \leq \widehat{\mathcal{L}}_M(f) + \frac{12VD^2}{\sqrt{M\left|\mathcal{M}_m\right|}} + 9VD^2\sqrt{\frac{\log\frac{1}{\delta}}{2M\left|\mathcal{M}_m\right|}},$$

where M is the number of multi-view clients and $|\mathcal{M}_m|$ is the number of samples in each multi-view client. Additionally, we have

$$\frac{1}{M}\sum_{m=1}^{M} d_{\mathcal{F}}\left(\tilde{\mathcal{D}}_v, \tilde{\mathcal{D}}\right) = d_{\mathcal{F}}\left(\tilde{\mathcal{D}}_v, \tilde{\mathcal{D}}\right) = \frac{2}{M}\sum_{m=1}^{M}\sup_{f\in\mathcal{F}}\left|\Pr_{\tilde{\mathcal{D}}_v}[f] - \Pr_{\tilde{\mathcal{D}}}[f]\right|.$$

According to Lemma 2 and Lemma 3, with probability at least $1-\delta$ for any $f \in \mathcal{F}$, we have

$$\begin{aligned} L_{S^v}(f) \leq & \widehat{L}_{S^v}(f) + \frac{10D^2}{\sqrt{S^v\left|\mathcal{S}_p\right|}} + 8D^2\sqrt{\frac{\log\frac{1}{\delta}}{2S^v\left|\mathcal{S}_p\right|}} \\ & + \sqrt{\frac{4}{M\left|\mathcal{M}_m\right|}\left(d\log\frac{2eM\left|\mathcal{M}_m\right|}{d} + \log\frac{4}{\delta}\right)} + d_{\mathcal{F}}\left(\tilde{\mathcal{D}}_v, \tilde{\mathcal{D}}\right) + \lambda_v, \end{aligned}$$

where S^v is the number of single-view clients with v-th view type and $|\mathcal{S}_p|$ is the number of samples in each single-view client.

D Experimental Settings

D.1 Datasets

We conduct the experiments on the following public datasets. **MNIST-USPS** [34] is a widely-used dataset for handwritten digits (0-9) and consists of 5000 examples with two views of digital images. The MNIST feature size is 28 × 28, while the USPS feature size is 16 × 16. **BDGP** [4] comprises 2500 examples related to drosophila embryos, each represented by a 1750-dimensional visual feature and a 79-dimensional textual feature. **Multi-Fashion** [41] is an image dataset featuring products like Coats, Dresses, and T-shirts, with images sized at 28 × 28. Following the approach in [10], which constructs a three-view version using 30,000 images, each instance includes three different images belonging to the same class. Consequently, the three views of each instance represent the same product with three distinct styles. **NUSWIDE** [9] is a web image dataset offering multiple views, including 65-dimensional color histogram, 226-dimensional block-wise color moments, 145-dimensional color correlogram, 74-dimensional edge direction histogram, and 129-dimensional wavelet texture. We utilize a total of 5000 samples for the evaluation of our proposed method.

Table 4: The statistics of experimental datasets.

Dataset	Clients	Sample	View	Class	Dimension
MNIST-USPS	24	5000	2	10	[[28, 28], [16,16]]
BDGP	12	2500	2	5	[1750,79]
Multi-Fashion	48	10000	3	10	[784, 784, 784]
NUSWIDE	24	5000	5	5	[65, 226, 145, 74, 129]

D.2 Implementation Details

The models of all methods are implemented on the PyTorch [33] platform using NVIDIA RTX-3090 GPUs. For an encoder-decoder pair, the encoder structure follows Input- $\mathrm{Fc}_{500} - \mathrm{Fc}_{500} - \mathrm{Fc}_{2000} - \mathrm{Fc}_{20}$, and the decoder is symmetric to the encoder. The non-linear mappings $\left\{\mathcal{H}^1\left(\mathbf{Z}^1;\Psi^1\right),\ldots,\mathcal{H}^V\left(\mathbf{Z}^V;\Psi^V\right)\right\}$ and $\mathcal{H}\left(\mathbf{Z};\Psi\right)$ adopt network architectures of $\mathrm{Fc}_{20} - \mathrm{Fc}_{256} - \mathrm{Fc}_{20}$ and $\mathrm{Fc}_{20V} - \mathrm{Fc}_{256} - \mathrm{Fc}_{20}$, respectively. The activation function is ReLU [12], and the optimizer uses Adam. For all the datasets used, the learning rate is fixed at 0.0003, the batch size is set to 256, and the temperature parameters τ_m and τ_p are both set to 0.5. Local pre-training is performed for 250

https://doi.org/10.52202/079017-1168

epochs on all datasets. After each communication round between the server and clients, local training is conducted for 10 epochs for the BDGP dataset and 25 epochs for other datasets on each client. The communication rounds between the server and clients are set to $R = 5$. All experiments in the paper involving FMCSC that are not mentioned are performed when M/S = 1:1.

We report some results that number of parameters and runtime by FMCSC to give the reader some information about the computational resources used by the method. Table 5 shows that the number of parameters and runtime by FMCSC are small and easy to reproduce.

Table 5: Number of parameters and runtime by FMCSC.

Dataset	MNIST-USPS	BDGP	Multi-Fashion	NUSWIDE
Number of parameters per client	3.4M-6.7M	3.5M-7M	3.4M-10.1M	2.8M-13.7M
Runtime	763.8s	108.5s	1183.5s	954.7s

D.3 Comparison Methods

We select 9 state-of-the-art methods, including HCP-IMSC [24], IMVC-CBG [39], DSIMVC [37], LSIMVC [27], ProImp [22], JPLTD [29], CPSPAN [18], FedDMVC [8] and FCUIF [36]. Among them, apart from FedDMVC and FCUIF, which are FedMVC methods, all the other comparison methods are centralized incomplete multi-view clustering methods. To ensure fair comparisons, we simplify the heterogeneous hybrid views scenario in our paper into a hybrid views scenario. Specifically, we concatenate the data distributed among the clients and use them as input for centralized methods, as shown in Figure 5. Among these, the data from multi-view clients can be considered

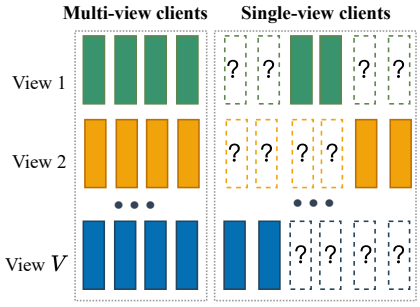


Figure 5: Comparison strategies.

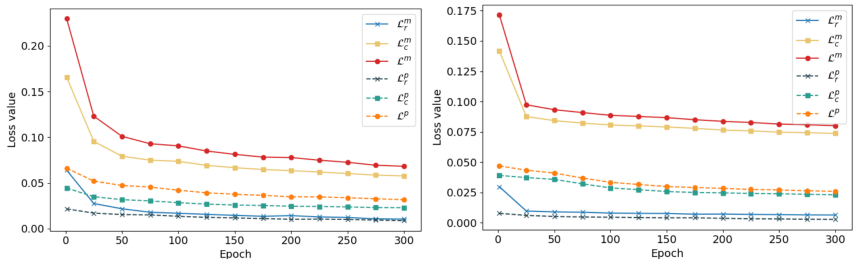
complete data, while the data from single-view clients can be regarded as missing data. It's worth noting that in our reported results, our method operates under the heterogeneous hybrid views scenario, whereas the other comparative methods operate under the hybrid views scenario. Although existing solutions can bypass the challenge of heterogeneous clients by simply concatenating data, the exposure of raw data, due to privacy concerns, may cause more data owners to refuse to participate in collaborative training. In contrast, our method can extract complementary clustering structures across clients without exposing their raw data, offering better privacy protection and performance improvement than current state-of-the-art methods.

E Additional Experiment Results

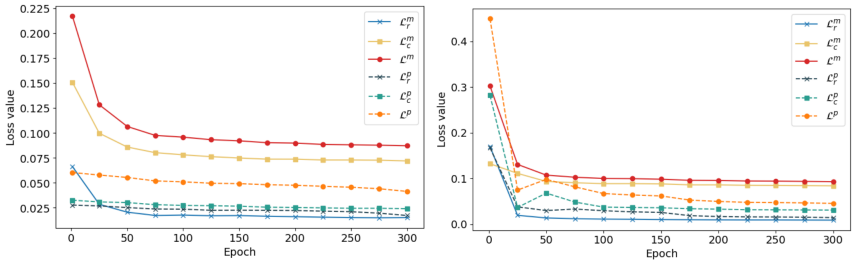
E.1 Convergence Analysis

Convergence analysis of the reconstruction loss, consistency loss, and total loss for multi-view clients and single-view clients is conducted using MNIST-USPS, BDGP, Multi-Fashion, and NUSWIDE. As illustrated in Figure 6, the increasing number of epochs leads to a gradual convergence of all loss functions, ultimately reaching a stable state. This clear observation serves as compelling evidence for the stability and effectiveness of our proposed model.

https://doi.org/10.52202/079017-1168



(a) MNIST-USPS (b) BDGP

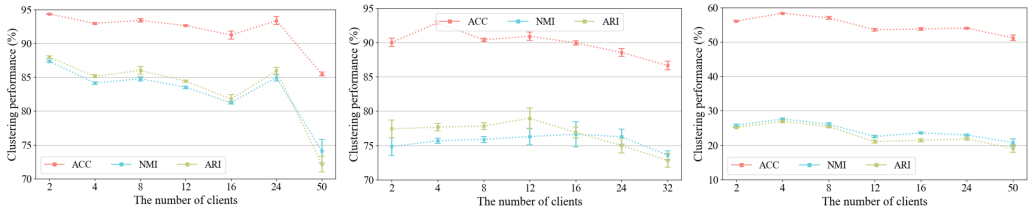


(c) Multi-Fashion (d) NUSWIDE

Figure 6: Convergence analysis on four datasets.

E.2 Attributes of Federated Learning

In this section, we present additional experimental results of FMCSC in various federated learning scenarios, including the number of clients and privacy. Figure 7 presents the impact of the number of clients on clustering performance for MNIST-USPS, BDGP, and NUSWIDE. It is observed that, with an increasing number of clients, the performance of FMCSC shows a slight decline but remains generally stable. However, on MNIST-USPS, as the number of clients reaches 50, the clustering performance experiences an unavoidable decrease due to the insufficient number of samples for each client.



(a) MNIST-USPS (b) BDGP (c) NUSWIDE

Figure 7: Scalability with the number of clients.

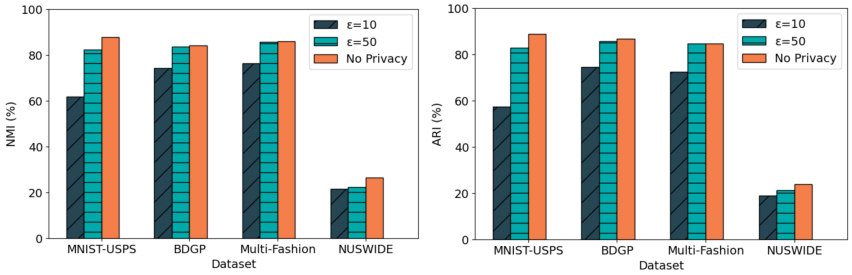


Figure 8: Sensitivity under privacy constraints when M/S = 2:1.

https://doi.org/10.52202/079017-1168

In FMCSC, we assume that all participating parties are semi-honest and do not collude with each other. An attacker faithfully executes the training protocol but may launch privacy attacks to infer the private data of other parties. Previous studies have demonstrated that sharing gradients can leak information about raw data [11]. To address this concern, we utilize differential privacy to further enhance the privacy guarantee of FMCSC. Differential privacy algorithms aim to introduce noise during individual data processing to protect against the disclosure of private information [1]. Formally, let $\mathcal{A}$ be a random mechanism that takes dataset D as input and belongs to set S. Assuming D_1 and D_2 are two neighboring datasets differing in only one point, $\mathcal{A}$ is (ε, δ) differentially private if:

$$\Pr\left[\mathcal{A}\left(\mathcal{D}_1\right) \in \mathcal{S}\right] \leq \exp(\varepsilon) \cdot \Pr\left[\mathcal{A}\left(\mathcal{D}_2\right) \in \mathcal{S}\right] + \delta \tag{13}$$

Here, ε and δ quantify the individual impact on the overall output in differential privacy. ε represents the strength of differential privacy, with smaller values indicating stronger privacy protection at the potential cost of increased noise. $\delta \sim \mathcal{O}\left(\frac{1}{\text{number of samples}}\right)$ is used to handle probabilistic exceptional cases. We implement Laplace noise to achieve differential privacy, and Figure 8 reports the clustering performance of FMCSC under different privacy strengths ε. Our observations reveal that FMCSC achieves both high performance and privacy at $\varepsilon = 50$, especially on the BDGP and Multi-Fashion datasets. However, as the level of noise increases at $\varepsilon = 10$, the performance of FMCSC unavoidably degrades.

F Broader Impacts

Heterogeneous hybrid views are prevalent in real-world scenarios. Our proposed method extends the application domain of existing FedMVC approaches to address complex scenarios such as healthcare and the Internet of Things (IoT). For example, hospitals in metropolitan areas use CT, X-ray, and EHR for disease detection, whereas remote areas often rely on a single detection method. Similarly, smartphones can capture both audio and images simultaneously, while recording devices are limited to collecting audio data only. Furthermore, this research is not expected to introduce any new negative societal impacts beyond those already known.

G Limitations

Our model addresses heterogeneous hybrid views in FedMVC, but it idealistically categorizes clients into two types: single-view clients and multi-view clients. In more realistic scenarios, multi-view clients can be further classified into full-view clients and partial-view clients based on the number of view types they have. Such a detailed categorization could encourage more heterogeneous devices to participate in federated learning, thereby enhancing the model's generalization ability and accelerating its application in fields like healthcare and finance. We will continue to explore this problem in future work and apply our findings to real-world scenarios.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we first introduce the relevant concepts of FedMVC and summarize existing methods, analyzing their shortcomings. Notably, existing methods assume isomorphic clients, prompting us to propose the concept of heterogeneous hybrid views. We focus on addressing the client gap and view gap that arise in this scenario. Our proposed method, which designs local-synergistic contrastive learning and global-specific weighting aggregation, aims to bridge these gaps and explore cluster structures. Additionally, we analyze the method's generalization performance and its effectiveness across various federated learning scenarios through theoretical and experimental evaluation.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Appendix G.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

https://doi.org/10.52202/079017-1168

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Please see Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please refer to Appendix D.2 for information on reproducing the main experimental results of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have uploaded the code for the new proposed method and the datasets used in the experiments in the Supplementary Material. The baseline methods compared in the paper are open source and can be reproduced directly after following the comparison strategy mentioned in Appendix D.3.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Appendix D for more experimental setting/details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report error bars in Table 1 and Figure 4 (b), which are standard deviations of the results obtained after five independent runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/079017-1168

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Appendix D.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the NeurIPS Code of Ethics and confirm that the paper complies.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see Appendix F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cite the original paper that produced datasets see Appendix D.1.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

https://doi.org/10.52202/079017-1168

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.