Bridging Inter-View and Client Heterogeneity: Federated Multi-View Clustering under Non-IID Data

Jiazhen Wang[1], **Xinyue Chen**[1], **Shuaiyu Liu**[1], **Zican He**[1], **Yazhou Ren**[1,2*], **Yi Wang**[3], **Shuyin Xia**[4,5]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
[3]Chongqing Ant Consumer Finance Co., Ltd
[4]Chongqing University of Posts and Telecommunications
[5]Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education, China
jiazhenwang0522@163.com, martinachen2580@gmail.com, 1827939309@qq.com, 2795291913@qq.com, yazhou.ren@uestc.edu.cn, haonan.wy@myxiaojin.cn, xiasy@cqupt.edu.cn

Abstract

Federated multi-view clustering (FedMVC) has been widely used to discover latent structures in distributed multi-view data, yet most methods still rely on the restrictive assumption of independent and identically distributed (IID) data. In practice, non-IID distributions with partial and imbalanced categories cause clients to learn biased and local representations, leading to model bias and unstable federated training performance. To address these issues, we propose a FedMVC framework called Bridging Inter-View and Client Heterogeneity: Federated Multi-View Clustering under Non-IID Data (H^2-FedMVC), which tackles both inter-view and client heterogeneity in mixed-view settings. We propose a leader-guided learning mechanism to address inter-view heterogeneity by enhancing complementary and discriminative feature learning, and an expert adaptive aggregation strategy to handle client heterogeneity by prioritizing well-matched experts in global updates. Theoretical analysis and experiments demonstrate superior performance in heterogeneous mixed-view FedMVC settings, with code provided at https://github.com/JiazhenWang-0522/H2-FedMVC.

1 Introduction

With the rapid advancement of sensor technologies and the internet [Driess *et al.*, 2023; Fei *et al.*, 2022; Wang *et al.*, 2025a; Zhang *et al.*, 2025], a multitude of distributed clients are now capable of collecting unlabeled data from multiple perspectives or modalities [Cui *et al.*, 2023]. While fully leveraging these data, preserving data privacy among clients has become a critical requirement, giving rise to Federated Multi-View Clustering (FedMVC [Chen *et al.*, 2023])—an emerging research direction that enables multiple clients to collaboratively learn a unified clustering model without sharing raw data. Such models can be widely applied in real-world scenarios such as recommendation systems and medical diagnosis [Che *et al.*, 2022], thus attracting increasing attention [Chen *et al.*, 2025a; Chen *et al.*, 2025b; Qin *et al.*, 2025; Feng *et al.*, 2025a; Ren *et al.*, 2026; Ren *et al.*, 2025].

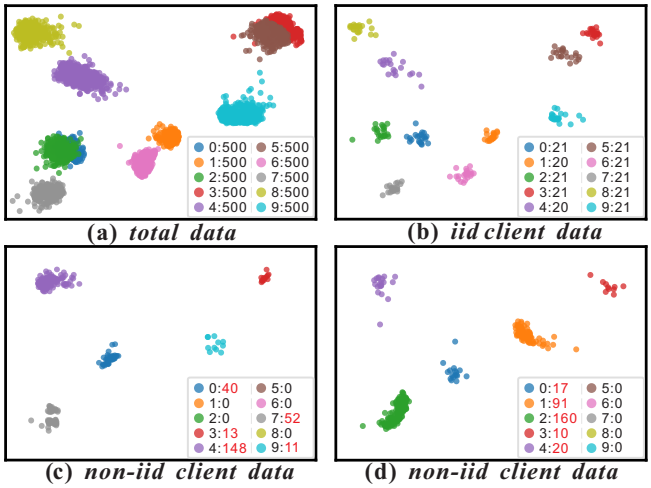

Figure 1: This figure shows the data distribution of the MNIST-USPS dataset under different conditions.

Existing FedMVC methods rely on idealized assumptions such as client homogeneity and independently and identically distributed (IID) data—all clients are either single-view or multi-view, with complete category coverage and balanced sample distributions. HFMVC [Jiang *et al.*, 2024] assumes a "single-view, multi-client" setup: each client has data from one view, possibly shared across clients, with roughly balanced sample sizes but differing category compositions. In contrast, FMCSC [Chen *et al.*, 2024] uses a "multi-view, multi-client" configuration: each client holds multi-view features of non-overlapping samples, with balanced sizes and full category coverage (Figure 1(b)). While these methods partially address non-IID data, they operate under mild heterogeneity and are unsuitable for more complex real-world settings. In challenging scenarios, clients may have access to either a single view or multiple views and are subject to multiple non-IID factors, such as incomplete category coverage,

*Corresponding author.

significant disparities in sample size, and severe class imbalance (Figure 1(c)–(d)). These compounded heterogeneities hinder effective alignment and integration of distributional differences, causing performance drops.

Unlike the aforementioned studies [Jiang *et al.*, 2024; Chen *et al.*, 2024], real-world FedMVC scenarios involve heterogeneous view structures and non-IID category distributions. Single-view and multi-view clients coexist, with inconsistent view availability across clients. Local data often has incomplete category coverage, severe class imbalance, or long-tail distributions. In healthcare federated learning, for instance [Hudaib *et al.*, 2025; Johnson *et al.*, 2016; Esteva *et al.*, 2019; Sheller *et al.*, 2018], hospitals typically focus on distinct clinical tasks such as cancer, Alzheimer's, or COVID-19 and their datasets exhibit significant class imbalance, limited category coverage, and large sample size variations, all impairing model convergence and generalization.

The presence of heterogeneous mixed-view data under non-IID constraints limits the applicability of many conventional FedMVC methods. These challenges can be broken down into two core issues: **(a) Inter-view heterogeneity**: Views collected by different clients vary significantly in terms of information type, quality, and expressive capability. For example, images typically contain rich visual structural information, while text emphasizes semantic expression, making it particularly challenging to effectively extract and fuse complementary information from heterogeneous views. **(b) Client heterogeneity** (non-IID issues): Different clients exhibit high inconsistency in category coverage and sample size (e.g., specialized hospitals in medical scenarios possess markedly different case types and volumes [Hudaib *et al.*, 2025]), posing serious challenges to cross-client collaborative modeling and joint optimization.

To address these challenges, this paper proposes a novel method termed **Bridging Inter-View and Client Heterogeneity: Federated Multi-View Clustering under Non-IID Data ($\mathbf{H}^2$-FedMVC)**, which aims to uniformly leverage single-view and multi-view data from heterogeneous clients to uncover clustering structures under mixed-view settings. As illustrated in Figure 2, we first observe that in the absence of unified supervision, directly aggregating local models can easily lead to model misalignment, while reliable cluster centers are difficult to establish at early training stages. To this end, we design a **cross-client consensus pretraining mechanism** to align local models and determine reliable cluster centers for each sample, thereby effectively alleviating model bias. Second, to handle view discreipancies, we introduce a **leader-guided learning mechanism**. At the individual level, the cluster center associated with each sample acts as a "leader" to guide contrastive updates; at the social level, clients learn from structurally similar and better-performing peer models to enable knowledge transfer. Third, to address the client non-IID issue, we propose an **expert adaptive aggregation strategy**, which dynamically adjusts aggregation weights based on client importance and model performance, enhancing accuracy while improving the stability and lightweight nature of the global model. These learning and aggregation strategies work collaboratively to significantly improve clustering performance in complex heterogeneous environments.

The main contributions of this paper are as follows:

1. A FedMVC framework is proposed to address view heterogeneity and non-IID issues in mixed-view scenarios.
2. A leader-guided learning mechanism and an expert-adaptive aggregation strategy are designed to address view and client heterogeneity, marking the first application of Mixture of Experts (MoE) to FedMVC.
3. The superior performance of H^2-FedMVC is verified in various federated scenarios. Experimental results demonstrate its excellent performance under ideal conditions as well as its robust and leading clustering performance even under complex non-ideal conditions.

2 Related Work

(1) Homogeneous FedMVC. This setting assumes all clients have the same view configuration, such as all being single-view or all being multi-view. This alignment enables consistent model structures and representation spaces, supporting parameter aggregation or representation alignment. Methods like HFMVC [Jiang *et al.*, 2024] further introduce a heterogeneity-aware mechanism to adaptively assess data distribution differences and extract cross-client consistency and complementarity via contrastive learning under vertical FedMVC [Chen *et al.*, 2024].

(2) Heterogeneous FedMVC. This setting allows single-view clients and multi-view clients to coexist, thus being closer to real application scenarios [Tian and Bennis, 2025]. FMCSC [Chen *et al.*, 2024] aims to integrate local and global learning via mutual information in federated settings, address cross-client and cross-view discrepancies, and robustly mine clustering structures from heterogeneous multi-view data.

Although existing methods have achieved certain results, the real FedMVC scenarios still face the challenges of heterogeneous mixed views, including the non-independent and identically distributed category distribution of clients, the structural heterogeneity of the coexistence of single-view and multi-view clients, and the uncertainty of the number and quality of views during the training process. To address the above complex scenarios, the H^2-FedMVC proposed in this paper can be regarded as a more general variant of heterogeneous FedMVC, which can handle the above complex scenarios by bridging the differences between clients and the gaps between views.

3 Methodology

3.1 Problem Definition

We consider a federated learning scenario where the data are non-IID. Suppose the dataset is $\{(\mathbf{x}_i^1, \mathbf{x}_i^2, \ldots, \mathbf{x}_i^V, y_i) \mid \mathbf{x}_i^v \in \mathbf{R}^{D_v},\ y_i \in \{1, \ldots, K\}\}_{i=1}^N$, where V is the number of views, K is the number of classes, and N is the number of samples. Here, $\mathbf{x}_i^v$ denotes the v-th view of the i-th sample, $\mathbf{R}^{D_v}$ is the feature space with dimensionality D_v for that view, and y_i is the class label of the i-th sample.

For client α, its local dataset is $D_\alpha = \{(\mathbf{x}_{\alpha_i}^1, \mathbf{x}_{\alpha_i}^2, \ldots, \mathbf{x}_{\alpha_i}^V, y_{\alpha_i}) \mid \mathbf{x}_{\alpha_i}^v \in \mathbf{R}^{D_v},\ y_{\alpha_i} \in C_\alpha\}_{i=1}^{N_\alpha}$, where $C_\alpha \subseteq \{1, \ldots, K\}$

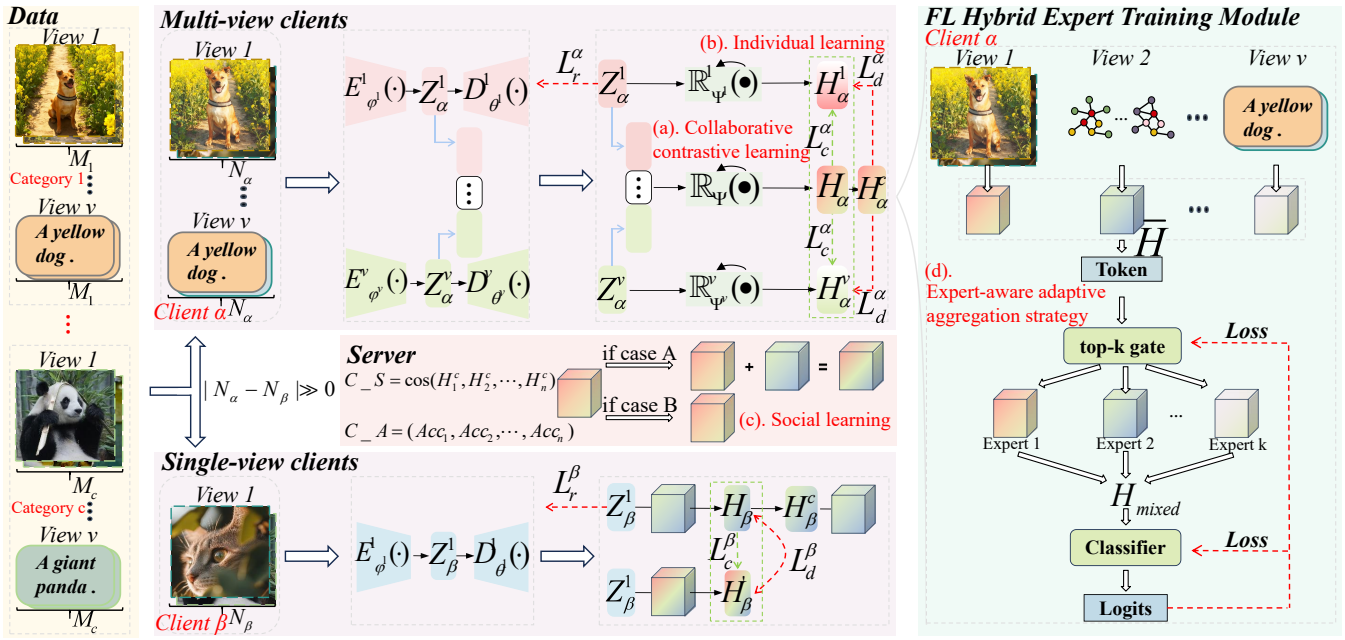

Figure 2: The framework of H^2-FedMVC operates as follows. First, each client performs (a) Collaborative contrastive learning (Section 3.2). Then, each client performs (b) Individual learning (Section 3.3). Subsequently, clients upload their expert models and parameters to the server to initiate Social learning (Section 3.3). The server executes an expert model update strategy and distributes the expert models to all clients (Section 3.4). Finally, locally, an Expert-Aware Adaptive Aggregation Strategy is implemented to generate the final expert model.

and $|C_\alpha| = c$, indicating that client α contains only c classes out of the K total classes. Let $n_{\alpha,k} = |\{\mathbf{x}_{\alpha_i} \mid y_{\alpha_i} = k\}|$, $k \in C_\alpha$ denote the number of samples in client α belonging to class k, and let $N_\alpha = \sum_{k \in C_\alpha} n_{\alpha,k}$ be the total number of samples on client α. For any $k_1, k_2 \in C_\alpha$, we have $|n_{\alpha,k_1} - n_{\alpha,k_2}| \gg 0$, and for another client β, $|N_\alpha - N_\beta| \gg 0$.

Assume there are in total $M + P$ clients, including M multi-view clients and P single-view clients. We train $\{f^1(\cdot;\mathbf{w}^1), \ldots, f^M(\cdot;\mathbf{w}^M), \ldots, f^{M+P}(\cdot;\mathbf{w}^{M+P})\}$ and mine complementary clustering structures. Here, $f^\alpha(\cdot;\mathbf{w}^\alpha)$ denotes the expert model produced by client α.

3.2 Cross-Client Consensus Pre-training

Multi-view datasets typically contain redundancy and random noise. Most existing methods employ self-supervised auto-encoder models, such as the auto-encoder (AE [Hinton and Zemel, 1993]) and variational auto-encoder (VAE [Kingma and Welling, 2013]), to learn high-level representations from raw data. In H^2-FedMVC, we adopt an encoder–decoder pair for each view, denoted as $E^v_{\phi^v}(\cdot)$ and $D^v_{\theta^v}(\cdot)$, with parameters ϕ^v and θ^v, respectively. For a multi-view client α, we assume its dataset size to be $|\mathcal{M}_\alpha|$. We define $\mathbf{z}^v_i = E^v_{\phi^v}(\mathbf{x}^v_i)$ as the latent representation of the v-th view of the i-th sample, and the auto-encoder output as $\hat{\mathbf{x}}^v_i = D^v_{\theta^v}(\mathbf{z}^v_i) \in \mathbf{R}^{D_v}$. The reconstruction loss is computed between the input $\mathbf{x}^v_i$ and the reconstructed output $\hat{\mathbf{x}}^v_i$. Furthermore, the local model can be pre-trained by minimizing the following objective:

$$L^\alpha_r = \frac{1}{|\mathcal{M}_\alpha|} \sum_{v=1}^{V} \sum_{i=1}^{|\mathcal{M}_\alpha|} \left\| \mathbf{x}^v_i - D^v_{\theta^v}\left(E^v_{\phi^v}(\mathbf{x}^v_i)\right) \right\|_2^2. \quad (1)$$

Similarly, for a single-view client p, we assume its dataset size to be $|\mathcal{S}_\beta|$; thus, it suffices to construct a single encoder–decoder pair. The pre-training objective L^β_r can adopt the same form as defined in Eq. (1).

$$L^\beta_r = \frac{1}{|\mathcal{S}_\beta|} \sum_{i=1}^{|\mathcal{S}_\beta|} \left\| \mathbf{x}^v_i - D^v_{\theta^v}\left(E^v_{\phi^v}(\mathbf{x}^v_i)\right) \right\|_2^2. \quad (2)$$

Meanwhile, since in the early stage of training, each expert model has difficulty in finding the corresponding cluster center for each cluster. Inspired by previous works of MVC [Wu *et al.*, 2024; Xu *et al.*, 2023; Xu *et al.*, 2022; Chen *et al.*, 2024], we introduce the collaborative contrastive learning to optimize the clustering structure of each client. It takes the low-level features obtained by the auto-encoder $\{(\mathbf{z}^1_i, \mathbf{z}^2_i, \ldots, \mathbf{z}^V_i)\}_{i=1}^{|\mathcal{M}_\alpha|}$ and obtains high-level features $\{(\mathbf{h}^1_i, \mathbf{h}^2_i, \ldots, \mathbf{h}^V_i)\}_{i=1}^{|\mathcal{M}_\alpha|}$ by stacking V nonlinear mappings $\{\mathcal{H}^1(\mathbf{Z}^1;\Psi^1), \ldots, \mathcal{H}^V(\mathbf{Z}^V;\Psi^V)\}$. In addition, a nonlinear mapping $\mathcal{H}(\mathbf{Z};\Psi) : \mathbf{R}^{\sum_{v=1}^V d_v} \to \mathbf{R}^d$ is used to construct the final representation,

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

where $\{\mathbf{h}_i\}_{i=1}^{|\mathcal{M}_\alpha|} = \mathbf{H} \in \mathbf{R}^{|\mathcal{M}_\alpha| \times d}$, and $\mathbf{Z} \in \mathbf{R}^{|\mathcal{M}_\alpha| \times \sum_{v=1}^V d_v}$. 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 $\{\mathbf{h}_i, \mathbf{h}_j^v\}_{j=i}^{v=1,\dots,V}$ as V positive feature pairs, and the remaining $\{\mathbf{h}_i, \mathbf{h}_j^v\}_{j\neq i}^{v=1,\dots,V}$ are $V(|\mathcal{M}_\alpha|-1)$ negative feature pairs. Then, we use cosine similarity to measure the similarity of feature pairs:

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

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

$$\mathcal{L}_c^\alpha = -\frac{1}{|\mathcal{M}_\alpha|} \sum_{v=1}^{V} \sum_{i=1}^{|\mathcal{M}_\alpha|} \log \frac{e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j^v)/\tau_\alpha}}{\sum_{j\neq i} e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_j^v)/\tau_\alpha}}. \tag{5}$$

Moreover, multi-view clients aim to assist single-view clients in bridging client gaps and discarding view-private information detrimental to clustering. Concretely,

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

where $f_\alpha^v(\cdot; \mathbf{w}_\alpha^v)$ represents the local model that can handle the v-th type of view, which is initialized by the global model $f_g^v(\cdot; \mathbf{w}^v)$. $f_\alpha(\cdot; \mathbf{w}_\alpha)$ represents the local model of multi-view client α.

For single-view client β, where each sample has only one view, we design model contrastive learning to achieve consistency. Specifically, client β contains data of the v-th view type $\{\mathbf{x}_i^v\}_{i=1}^{|\mathcal{S}_\beta|}$, with its local model as $f_\beta(\cdot; \mathbf{w}^v)$. To capture common semantics, we use the same approach as for multi-view clients: two non-linear mappings, $\mathcal{H}^v(\mathbf{Z}^v; \Psi^v)$ and $\mathcal{H}(\mathbf{Z}; \Psi)$, generate high-level features $\mathbf{H}^v$ and common semantics $\mathbf{H}$. The global model $f_g(\cdot; \mathbf{w}^v)$ after aggregation further enhances its ability to learn generalized common semantics. To align the local model with the global model, we define the model contrastive loss:

$$\mathcal{L}_c^\beta = -\frac{1}{|\mathcal{S}_\beta|} \sum_{i=1}^{|\mathcal{S}_\beta|} \log \frac{e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^g)/\tau_\beta}}{e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{h}_i^g)/\tau_\beta} + e^{\mathrm{sim}(\mathbf{h}_i, \mathbf{z}_i^v)/\tau_\beta}}, \tag{7}$$

where $\{\mathbf{h}_i^g\}_{i=1}^{|\mathcal{S}_\beta|}$ represents the output of the local data after being processed by the global model, and τ_β denotes the temperature parameter.

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

$$\mathcal{L}^\alpha = \mathcal{L}_r^\alpha + \mathcal{L}_c^\alpha, \qquad \mathcal{L}^\beta = \mathcal{L}_r^\beta + \mathcal{L}_c^\beta, \tag{8}$$

where $\mathcal{L}_r^\alpha$ and $\mathcal{L}_r^\beta$ are used as reconstruction losses to learn representations for each view individually. Meanwhile, $\mathcal{L}_c^\alpha$ and $\mathcal{L}_c^\beta$ are employed to discover common semantics across views, facilitating the exploration of complementary clustering structures across clients.

3.3 Leadership-Guided Learning Mechanism

The heterogeneities between views can be divided into *intra-client view heterogeneities* and *inter-client view heterogeneities*. For this purpose, inspired by the particle swarm optimization (PSO [Kennedy and Eberhart, 1995]), we have designed the Client-level Leadership-Guided Learning Mechanism (Individual learning) and the Model-level Leadership-Guided Learning Mechanism (Social learning).

Client-level Leadership-Guided Learning Mechanism

An ideal clustering structure requires the maximum intra-cluster distance to be smaller than the minimum inter-cluster distance [Bengio *et al.*, 2013]. In such cases, samples within the same cluster naturally converge toward their centroid. For samples of the same category [Kulis, 2013], the angular separation between their refined features h should be minimized, whereas for samples from different categories, it should be maximized.

From a collective view, a cluster can be seen as a population with its centroid acting as a "leader" that captures shared characteristics and serves as a prototype. However, most contrastive learning methods use individual samples as comparison units [Chen *et al.*, 2024; Jiang *et al.*, 2024]. Although these methods are effective in capturing discriminative features at the sample level, they often overlook the consistency at the group level and are unable to model group features.

For multi-view client α, samples of each category naturally converge toward their corresponding cluster center. To exploit this structure and mitigate cross-view distribution discrepancies, we introduce a Leader-Guided Learning Mechanism for instance-level feature learning. This mechanism uses cluster centroids as leaders, encouraging samples to approach their own category centroid while being repelled from those of other categories. The corresponding optimization objective is defined as:

$$\mathcal{J}_d^\alpha = \frac{1}{|\mathcal{N}_\alpha|} \sum_{v=1}^{V} \sum_{i=1}^{|n_{\alpha,k_a}|} \log \frac{e^{sim(\mathbf{h}_{k_a,c}^v, \mathbf{h}_{k_a,i}^v)/\tau_\alpha}}{\sum_{\substack{a\neq b \\ i\neq j}} e^{sim(\mathbf{h}_{k_a,c}^v, \mathbf{h}_{k_b,j}^v)/\tau_\alpha}}, \tag{9}$$

where $k_a, k_b \in \mathcal{C}_\alpha$, k_a and k_b denote the class labels of the k_a-th and k_b-th samples, respectively, and $|n_{\alpha,k_a}|$ represents the number of samples belonging to class k_a on client α. We define the objective function as $\mathcal{L}_d^\alpha = -\mathcal{J}_d^\alpha$ so as to minimize the loss $-\mathcal{J}_d^\alpha$.

Meanwhile, as the client-side models are continuously updated during the training process, in each round of aggregation or election, each group dynamically produces a new *"leader"*, namely the updated clustering center. The update rule [MacQueen, 1967] is given as:

$$\mathbf{h}_{k_a,c} = \frac{1}{|n_{\alpha,k_a}|} \sum_{\mathbf{h}_i \in n_{\alpha,k_a}} \mathbf{h}_i. \tag{10}$$

For single-view clients, due to their lack of complete view information, they mainly learn through the Model-level Leadership-Guided Learning Mechanism.

Model-level Leadership-Guided Learning Mechanism
In practical application scenarios, a single client usually cannot cover all categories of samples, but different clients often have intersections and overlaps in some categories. Due to sharing similar category data, the models trained by these clients often have high similarity in feature representation capabilities and network parameter spaces. Inspired by this phenomenon, we propose a model-level leadership-driven learning mechanism to achieve "social learning" among models. This mechanism promotes the effective dissemination and coevolution of knowledge among clients by guiding client models to converge towards a "leader model" with stronger representativeness or better performance.

To preserve privacy, for client α, only its expert models $f^1(\cdot; w^1), \ldots, f^{M+P}(\cdot; w^{M+P})$, together with their accuracies and the corresponding cluster centers $H_\alpha = \{(h_{\alpha,k_1}, \ldots, h_{\alpha,k_{|C_\alpha|}}) \mid k_1, \ldots, k_{|C_\alpha|} \in C_\alpha\}$, are transmitted to the server. To address the inter-client view differences, the server computes the similarity between client α and client γ as:

$$\begin{aligned}\mathrm{Sim}_{\alpha,\gamma} &= \sigma\big(\cos^2\langle H_\alpha, H_\gamma\rangle\big) \\ &= \sigma\left(\left(\sum_{k\in C_\alpha}\sum_{l\in C_\gamma}\frac{h_{\alpha,k}^\top h_{\gamma,l}}{\|h_{\alpha,k}\|_2\,\|h_{\gamma,l}\|_2}\right)^2\right).\end{aligned} \tag{11}$$

Assume that the client most similar to client α is client γ, that is $\gamma = \arg\max_{\zeta\neq\alpha}\mathrm{Sim}_{\alpha,\zeta}$. The model is then updated according to the following rule:

$$f^\alpha(\cdot; w^\alpha) = \begin{cases}(1-\ell)\, f^\alpha(\cdot) + \ell\, f^\gamma(\cdot), & \text{case A}\\ f^\alpha(\cdot), & \text{case B}\end{cases}, \tag{12}$$

where Case A corresponds to $Acc_\alpha < Acc_\gamma$, Case B corresponds to $Acc_\alpha \geq Acc_\gamma$, and ℓ denotes the learning rate.

Finally, each client applies the K-means [MacQueen, 1967] 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:

$$\min_{\mathbf{c}_1,\mathbf{c}_2,\ldots,\mathbf{c}_K}\sum_{i=1}^{|\mathcal{M}_m|}\sum_{j=1}^{K}\|\mathbf{h}_i - \mathbf{c}_j\|^2. \tag{13}$$

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.

3.4 Expert Adaptive Aggregation Strategy

Inspired by these articles [Shazeer *et al.*, 2017; Fedus *et al.*, 2022; Khuong *et al.*, 2025; Feng *et al.*, 2025b], we integrate MoE into FedMVC. After the local training of expert models is completed on each client, the trained expert models are uploaded to the server [McMahan *et al.*, 2017]. The server then aggregates all the received expert models and uniformly distributes them back to each client. To prevent user privacy leakage, the training of the MoE network is carried out entirely locally on the client side. At this stage, only the model parameters of the MoE network are transferred between clients, without sharing any raw data or user sensitive information.

For the α-th client, the local dataset is defined as $\mathcal{D}_\alpha = \big\{(\mathbf{x}_{\alpha i}^1, \ldots, \mathbf{x}_{\alpha i}^V; y_{\alpha i}) \mid \mathbf{x}_{\alpha i}^v \in \mathbf{R}^p,\ y_{\alpha i} \in \mathcal{C}_\alpha\big\}_{i=1}^{N_\alpha}$. Each input sample $\mathbf{x}_{\alpha i}$ is processed by all expert models $\{f^1(\cdot; w^1), \ldots, f^{M+P}(\cdot; w^{M+P})\}$, producing a set of refined features $\{H_{\alpha i}^1, H_{\alpha i}^2, \ldots, H_{\alpha i}^{M+P}\}$. These refined features are averaged to obtain $\bar{H}_{\alpha i}$, and the routing network applies layer normalization to generate an embedding token [Riquelme *et al.*, 2021], given by $x = \mathrm{LN}(\bar{H}_{\alpha i})$. Subsequently, a linear transformation computes logits over N experts as $\mathcal{G} = Wx + b$, where $\mathcal{G} \in \mathbf{R}^N$. The top-$K$ experts are selected based on these logits, denoted as $\{i_1, \ldots, i_K\} = \mathrm{TopK}(\mathcal{G})$, ensuring sparse activation of the most relevant experts.

The logits corresponding to the selected experts are passed through a Softmax function to compute their normalized weights,

$$\mathcal{W}_j = \frac{\exp(\mathcal{G}_{i_j})}{\sum_{m=1}^K \exp(\mathcal{G}_{i_m})}, \tag{14}$$

reflecting the relative importance of each expert.

A weighted combination of the selected expert outputs is then computed, yielding the mixed representation

$$h_{\text{mixed}} = \sum_{j=1}^{K}\mathcal{W}_j h_{i_j}. \tag{15}$$

Finally, the classification head, implemented as a learnable linear layer, generates the predicted label

$$\hat{y} = \arg\max\left(W_c h_{\text{mixed}} + b_c\right). \tag{16}$$

To finalize the training iteration, the gating network and classifier are jointly optimized using the cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{k=1}^{K} y_k \log\left(\frac{\exp(\text{logits}_k)}{\sum_q \exp(\text{logits}_q)}\right), \tag{17}$$

where K denotes the number of classes, logits_k represents the unnormalized prediction score for class k, and y_k is the one-hot encoded ground-truth label. The Softmax function converts logits into a normalized probability distribution over all classes. Cross-entropy loss penalizes low confidence in the true class and enables joint backpropagation-based optimization of the classifier and gating network in the MoE framework.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluated our method on four multi-view datasets. **MNIST-USPS** [Peng *et al.*, 2019] includes 5,000 samples from two handwritten digit image sources, forming two views. **BDGP** [Cai *et al.*, 2012] contains 2,500 samples across five fruit fly categories, each with text and visual views. **Multi-Fashion** [Xiao *et al.*, 2017] has 10,000 samples in 10 categories, where different styles of the same object are treated as three views. **NUSWIDE** [Chua *et al.*, 2009] comprises 5,000 images with five views. For specific client settings, please refer to the supplementary materials.

	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]	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
	DSIMVC [2022]	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
	ProImp [2023]	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
	FedDMVC [2023]	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]	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±0.2	64.4±0.9	53.5±0.8
	FMCSC [2024]	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
	Energy-DIMC [2025b]	95.1±0.5	88.6±1.0	89.5±0.9	90.6±1.3	80.5±2.2	80.6±2.4	86.4±1.4	73.2±2.2	72.6±2.4
	ours	**99.0±0.0**	**97.2±0.1**	**97.9±0.0**	**97.3±0.1**	**93.2±0.3**	**94.2±0.2**	**95.4±0.0**	**89.0±0.2**	**90.1±0.2**
BDGP	HCP-IMSC [2022]	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
	DSIMVC [2022]	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
	ProImp [2023]	91.6±0.3	82.4±3.8	80.0±0.9	90.4±1.5	62.0±1.6	79.3±1.6	75.6±0.5	52.2±0.0	46.1±0.1
	FedDMVC [2023]	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]	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 [2024]	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
	Energy-DIMC [2025b]	97.0±0.8	90.6±2.2	92.7±1.8	91.4±1.9	77.0±3.1	80.2±3.7	90.0±0.9	73.1±1.7	76.8±1.8
	ours	**97.6±0.0**	**92.2±0.1**	**94.2±0.0**	**96.7±0.3**	**89.2±1.8**	**91.9±1.4**	**93.3±0.9**	**80.9±1.9**	**84.4±2.2**
Multi-Fashion	HCP-IMSC [2022]	70.6±0.1	67.4±0.1	57.7±0.0	67.1±0.1	64.7±0.1	53.1±0.1	59.9±0.7	56.4±0.9	41.8±1.1
	DSIMVC [2022]	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
	ProImp [2023]	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
	FedDMVC [2023]	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]	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 [2024]	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±1.0	76.3±1.0
	Energy-DIMC [2025b]	89.7±0.2	83.0±0.3	80.3±0.3	86.2±0.4	77.8±0.2	77.4±0.4	82.9±0.4	73.8±0.7	69.2±0.7
	ours	**94.6±0.1**	**89.1±0.3**	**88.9±0.4**	**92.0±0.0**	**84.6±0.0**	**83.7±0.1**	**89.0±0.3**	**80.7±0.4**	**78.5±0.6**
NUSWIDE	HCP-IMSC [2022]	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
	DSIMVC [2022]	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
	ProImp [2023]	38.4±0.1	11.1±0.0	8.3±0.1	37.1±0.4	15.0±0.1	7.6±0.2	34.3±0.7	8.0±0.0	5.4±0.0
	FedDMVC [2023]	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±0.8	3.4±0.2
	FCUIF [2024]	45.2±0.3	15.0±0.3	14.1±0.2	42.0±0.5	10.0±0.6	9.2±0.5	38.2±0.4	9.6±0.3	8.2±0.3
	FMCSC [2024]	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
	Energy-DIMC [2025b]	57.1±0.4	29.6±0.6	25.6±0.4	52.2±0.5	25.5±0.6	24.3±0.5	54.0±0.5	22.1±0.6	21.1±0.6
	ours	**65.4±0.3**	**33.9±0.3**	**34.8±0.4**	**60.2±0.5**	**27.7±0.5**	**28.4±0.5**	**57.8±0.1**	**24.8±0.1**	**25.1±0.1**

Table 1: Clustering results (mean±std%) of all methods on four datasets, where each client contains all classes under the Dirichlet(∞) setting. 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	FMCSC(Dirichlet(∞)) [2024]	45.2±0.3	43.8±2	32.2±2	42.6±1.5	39.9±3.3	28.8±2.6	41.3±2	36.3±3.3	26.1±2.6
	FMCSC(Dirichlet(0.5)) [2024]	35.2±6.2	31.1±5	21±4.7	32.3±3.3	26.2±4.4	17.3±3.4	29.9±5.4	22±7.4	14.8±5.5
	Ours(Dirichlet(∞))	**97.4±0.2**	**93.3±0.4**	**94.3±0.5**	**94.7±0.5**	**88.2±0.9**	**88.7±0.9**	87.8±0.7	79.1±0.8	76.8±1.1
	Ours(Dirichlet(0.5))	94.4±1.8	87.8±3.2	88.2±3.7	89.6±2.6	79.8±4.4	78.5±5.2	**89.3±0.6**	78.5±0.8	**78.1±1.1**
BDGP	FMCSC(Dirichlet(∞)) [2024]	49.9±9.1	34.6±12.6	29.9±11.4	51.1±3.4	33.1±4.5	29.3±4.3	46.2±3.5	25.2±4.6	22±4.1
	FMCSC(Dirichlet(0.5)) [2024]	42.3±3.7	19.6±2.2	16.9±2.3	38.7±2.4	17.6±4.3	14.7±3.2	42.8±5.6	19.3±5.9	16.5±5.7
	Ours(Dirichlet(∞))	**95.5±1.0**	**87.4±2.4**	**89.4±2.3**	**93.9±2.8**	**83.6±6.1**	**85.8±6.0**	**91.0±1.5**	**76.5±3.0**	**79.2±3.1**
	Ours(Dirichlet(0.5))	93.2±2.3	81.4±5.0	84.1±5.0	86.8±4.0	68.6±7.5	70.9±7.8	82.0±5.1	59.7±8.8	61.7±9.3
Multi-Fashion	FMCSC(Dirichlet(∞)) [2024]	46.4±0.8	51.1±1.6	35.9±1.1	45.7±0.9	49.0±2.5	34.0±2.1	44.3±1.5	45.8±2.5	31.6±2.3
	FMCSC(Dirichlet(0.5)) [2024]	42.9±1.9	42.7±3.3	29.3±2.7	40.0±4.1	41.3±5.4	27.6±4.4	36.3±2.1	34.1±5.5	22±3.4
	Ours(Dirichlet(∞))	**91.9±0.4**	**85.2±0.7**	**83.7±0.8**	**88.1±0.9**	**79.4±0.9**	**76.7±1.4**	**82.8±1.8**	**73.4±1.7**	**68.0±3.1**
	Ours(Dirichlet(0.5))	85.0±3.5	78.7±3.1	72.6±5.3	80.7±2.4	72.8±2.7	66.1±3.8	72.7±4.7	63.4±4.2	54.8±5.8
NUSWIDE	FMCSC(Dirichlet(∞)) [2024]	41.7±3.5	15.7±3.1	14.6±3.3	39.8±2.5	13.8±1.4	12.6±1.9	36.4±1.8	9.9±1.9	9±1.6
	FMCSC(Dirichlet(0.5)) [2024]	35.1±2.5	9±2.3	8±2.1	33.4±3.1	7.6±2	6.6±1.8	32.6±2.5	6.7±1.2	6±1.3
	Ours(Dirichlet(∞))	**58.3±2.9**	**25.5±2.7**	**25.4±3.2**	**55.2±1.4**	**21.2±1.7**	**21.1±1.8**	**52.8±1.8**	**18.5±1.2**	**18.4±2.2**
	Ours(Dirichlet(0.5))	56.2±2.9	23.0±2.0	22.6±2.8	54.3±2.2	20.7±3.3	20.5±3.7	51.8±1.9	15.7±2.6	13.7±1.9

Table 2: Clustering results (mean±std%) of all methods on four datasets, where each client exclusively encompasses a subset of classes under the Dirichlet(0.5) and Dirichlet(∞) setting. The best and second best results are denoted in **bold** and underline.

	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		✓	✓	✓	45.1±3.6	38.1±1.3	24.7±2.2	63.6±7.8	55.0±9.1	48.2±32.3	77.8±0.4	67.7±1.9	56.3±1.1	34.2±0.2	5.5±0.3	5.0±0.4
Item-2	✓		✓	✓	81.7±0.6	66.6±1.1	64.6±1.6	93.1±0.3	80.3±2.4	83.8±1.8	83.5±0.2	73.6±0.4	69.7±0.4	43.1±0.9	10.2±0.6	9.6±0.5
Item-3	✓	✓		✓	96.6±0.0	92.7±0.0	93.8±0.0	95.7±0.0	86.7±0.6	89.7±0.3	91.4±0.1	84.5±0.1	83.6±0.2	46.8±35.4	13.7±35.3	13.5±38.6
Item-4	✓	✓	✓		95.0±0.3	88.7±0.9	89.4±1.1	91.5±4.1	76.8±18.1	80.1±19.6	89.0±0.0	81.1±0.0	78.9±0.0	48.9±24.6	15.5±31.1	15.2±33.9
Item-5	✓	✓	✓	✓	97.3±0.1	93.2±0.3	94.2±0.2	96.7±0.3	89.2±1.8	91.9±1.4	92.0±0.0	84.6±0.0	83.7±0.1	60.2±0.5	27.7±0.5	28.4±0.5

Table 3: Ablation study results (mean±std%) under different component combinations on four datasets.

Comparison Methods. We tested two heterogeneous settings to demonstrate our method's adaptability: Dirichlet(∞) and Dirichlet(0.5). Smaller Dirichlet($\cdot$) parameters indicate higher heterogeneity and more imbalanced class and label distributions. Under Dirichlet(∞), we compared seven recent methods: HCP-IMSC [Li *et al.*, 2022], DSIMVC [Tang and Liu, 2022], ProImp [Li *et al.*, 2023], FedDMVC [Chen *et al.*, 2023], FCUIF [Ren *et al.*, 2024], FMCSC [Chen *et al.*, 2024] and Energy-DIMC [Wang *et al.*, 2025b]. FedDMVC, FCUIF, and FMCSC are federated multi-view clustering methods; the others are centralized incomplete multi-view clustering approaches. For fair comparison, we concatenated data across all clients as input for centralized methods. Multi-view clients provided complete data; single-view clients were treated as having missing views.

Implementation Details. For each encoder–decoder pair, the encoder architecture is *Input – Fc500 – Fc500 – Fc2000 – Fc15*, and the decoder is symmetric to the encoder. We set the temperature parameters to $\tau_\alpha = 0.6$ and $\tau_\beta = 0.7$, and use a batch size of 256. The output dimension d is fixed to 15 for all models. Unless otherwise specified, all experiments involving H^2-FedMVC are conducted under the condition $M/S = 1:1$.

4.2 Results and Analysis

Clustering Results. Table 1 presents quantitative results under various heterogeneous mixed-view scenarios, with experiments conducted under the Dirichlet(∞) setting. Each experiment was repeated five times independently, and averages with standard deviations are reported. By varying the ratio of multi-view to single-view clients, we created different scenarios. Results show that H^2-FedMVC outperforms recent methods across all settings, achieving strong clustering performance while preserving data privacy. As the proportion of single-view clients increases, all methods degrade to some extent, as expected. Notably, even when single-view clients double the number of multi-view clients, H^2-FedMVC maintains excellent performance, demonstrating robustness.

Table 2 presents comparison results under the Dirichlet(∞) and Dirichlet(0.5) settings for heterogeneous mixed-view scenarios where no client contains all categories. Each experiment was repeated five times independently, and averages with standard deviations are reported. Different scenarios were created by adjusting the ratio of multi-view to single-view clients. Results show that H^2-FedMVC remains stable in this challenging setting: under Dirichlet(∞), it is almost superior to all the comparison methods. Under Dirichlet(0.5), except in the case where the dataset is MNIST-USPS and $M/S = 1:2$, it ranks second with no significant drop in accuracy.

Combining the results of the above two sets of experiments, it can be seen that the proposed H^2-FedMVC method exhibits excellent and stable performance in both balanced and unbalanced heterogeneous mixed-view scenarios, fully verifying its effectiveness and robustness.

Ablation Study. Components A and B correspond to the *consensus pre-training process* and the *cluster center initialization strategy*, respectively. Component C denotes the *Model-level Leadership-Driven Learning Mechanism*, which is formally defined in Eq. (12), while Component D represents the *Expert Adaptive Aggregation Strategy*.

As shown in Table 3, components A and B play a crucial role in clustering performance: the high-level features learned by component A directly affect the quality of subsequent refined features, thereby determining the separability of clustering; without this component, the model struggles to achieve effective clustering results. Component B is used to initialize clustering centers, and its removal leads to unstable clustering in the early training stage and a performance drop. After removing component C, both single-view and multi-view clients adopt a unified expert learning strategy. Experimental results show that the introduction of component C can significantly improve performance, verifying the effectiveness of the model-level comparison and guidance mechanism. Component D achieves a balance between performance and lightweight by aggregating the top-K expert models, ensuring category coverage while controlling the model size.

Other Analysis Results. Due to space limitations, temperature optimization and client-number analyses are placed in supplementary materials, showing our algorithm's robustness and stability under incomplete multi-view and non-IID data.

5 Conclusion

In this paper, we propose H^2-FedMVC, a novel method for real-world scenarios with both IID and non-IID heterogeneous mixed views, to mine latent clustering structures across multiple clients. We first design a cross-client consensus pre-training strategy to align local models while collaboratively optimizing clustering structures. Then, we introduce a leader-guided learning mechanism to reduce view heterogeneity and an expert adaptive aggregation strategy to address client heterogeneity. Both theoretical analysis and extensive experimental results demonstrate that H^2-FedMVC consistently outperforms state-of-the-art methods across various IID settings, while also maintaining robust and competitive performance under challenging non-IID scenarios.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (No. 62476052) and the Open Fund of the Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education (No. CBDIS202501).

References

[Bengio *et al.*, 2013] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *TPAMI*, pages 1798–1828, 2013.

[Cai *et al.*, 2012] Xiao Cai, Hua Wang, Heng Huang, and Chris Ding. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, pages i16–i24, 2012.

[Che *et al.*, 2022] 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. *ACM TIST*, pages 1–23, 2022.

[Chen *et al.*, 2023] 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.

[Chen *et al.*, 2024] Xinyue Chen, Yazhou Ren, Jie Xu, Fangfei Lin, Xiaorong Pu, and Yang Yang. Bridging gaps: Federated multi-view clustering in heterogeneous hybrid views. In *NeurIPS*, pages 37020–37049, 2024.

[Chen *et al.*, 2025a] Xinyue Chen, Jinfeng Peng, Yuhao Li, Xiaorong Pu, Yang Yang, and Yazhou Ren. An effective and secure federated multi-view clustering method with information-theoretic perspective. In *ICML*, pages 8871–8889, 2025.

[Chen *et al.*, 2025b] Zhe Chen, Xiao-Jun Wu, Tianyang Xu, Hui Li, and Josef Kittler. Multi-layer multi-level comprehensive learning for deep multi-view clustering. *Information Fusion*, page 102785, 2025.

[Chua *et al.*, 2009] 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*, page 1–9, 2009.

[Cui *et al.*, 2023] 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 44847–44859, 2023.

[Driess *et al.*, 2023] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[Esteva *et al.*, 2019] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, pages 24–29, 2019.

[Fedus *et al.*, 2022] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, pages 1–39, 2022.

[Fei *et al.*, 2022] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, Hao Sun, and Ji-Rong Wen. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, page 3094, 2022.

[Feng *et al.*, 2025a] Wei Feng, Zeyu Bi, Qianqian Wang, and Bo Dong. Federated multi-view graph clustering with incomplete attribute imputation. In *IJCAI*, pages 5118–5126, 2025.

[Feng *et al.*, 2025b] Yu Feng, Yangli-ao Geng, Yifan Zhu, Zongfu Han, Xie Yu, Kaiwen Xue, Haoran Luo, Mengyang Sun, Guangwei Zhang, and Meina Song. Pmmoe: Mixture of experts on private model parameters for personalized federated learning. In *ACM WWW*, pages 134–146, 2025.

[Hinton and Zemel, 1993] Geoffrey E. Hinton and Richard S. Zemel. Autoencoders, minimum description length and helmholtz free energy. In *NeurIPS*, page 3–10, 1993.

[Hudaib *et al.*, 2025] Amjad Hudaib, Nadim Obeid, Amjad Albashayreh, Hebah Mosleh, Yahya Tashtoush, and Georgi Hristov. Exploring the implementation of federated learning in healthcare: a comprehensive review. *Cluster Computing*, pages 1–21, 2025.

[Jiang *et al.*, 2024] Xiaorui Jiang, Zhongyi Ma, Yulin Fu, Yong Liao, and Pengyuan Zhou. Heterogeneity-aware federated deep multi-view clustering towards diverse feature representations. In *ACM MM*, page 9184–9193, 2024.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, pages 1–9, 2016.

[Kennedy and Eberhart, 1995] James Kennedy and Russell Eberhart. Particle swarm optimization. *ICNN*, pages 1942–1948, 1995.

[Khuong *et al.*, 2025] Duy Khuong, An D Nguyen, Duy Nguyen, and Kok-Seng Wong. Fedkoe: Enhancing federated multimodal learning through knowledge of experts. In *FL-AsiaCCS*, pages 1–6, 2025.

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

[Kulis, 2013] Brian Kulis. Metric learning: A survey. *Found. Trends Mach. Learn*, pages 287–364, 2013.

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

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

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

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. *AISTATS*, pages 1273–1282, 2017.

[Peng *et al.*, 2019] 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.

[Qin *et al.*, 2025] Lang Qin, Xiaorui Jiang, Yu Gao, and Yong Liao. Federated deep incomplete multi-view clustering with heterogeneity-matching and attention-based imputation. In *MMAsia*, pages 1–8, 2025.

[Ren *et al.*, 2024] 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*, page 108:102357, 2024.

[Ren *et al.*, 2025] Yazhou Ren, Junlong Ke, Zichen Wen, Tianyi Wu, Yang Yang, Xiaorong Pu, and Lifang He. Multi-view graph clustering via node-guided contrastive encoding. In *ICML*, pages 51421–51435, 2025.

[Ren *et al.*, 2026] Yazhou Ren, Zichen Wen, Junlong Ke, Chenhang Cui, Yonghao Huang, Xinyue Chen, Philip S. Yu, and Lifang He. Enhancing multi-view clustering: A sufficient information-theoretic approach for consistency acquisition and redundancy elimination. *TPAMI*, pages 1–16, 2026.

[Riquelme *et al.*, 2021] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. In *NeurIPS*, pages 8583–8595, 2021.

[Shazeer *et al.*, 2017] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, pages 1–19, 2017.

[Sheller *et al.*, 2018] Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. *BrainLes (MICCAI)*, pages 92–104, 2018.

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

[Tian and Bennis, 2025] Zhuojun Tian and Mehdi Bennis. Compositional distributed learning for multi-view perception: A maximal coding rate reduction perspective. *Signal Processing Letters*, pages 4409–4413, 2025.

[Wang *et al.*, 2025a] Shuai Wang, Malu Zhang, Dehao Zhang, Ammar Belatreche, Yichen Xiao, Yu Liang, Yimeng Shan, Qian Sun, Enqi Zhang, and Yang Yang. Spiking vision transformer with saccadic attention. In *ICLR*, pages 24148–24169, 2025.

[Wang *et al.*, 2025b] Ziyu Wang, Yiming Du, Rui Ning, and Lusi Li. Energy-based deep incomplete multi-view clustering. In *ACM MM*, pages 1686–1694, 2025.

[Wu *et al.*, 2024] 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. In *TMM*, pages 9150–9162, 2024.

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

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

[Xu *et al.*, 2023] Jie Xu, Shuo Chen, Yazhou Ren, Xiaoshuang Shi, Hengtao Shen, Gang Niu, and Xiaofeng Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. In *NeurIPS*, pages 1119–1131, 2023.

[Zhang *et al.*, 2025] Dehao Zhang, Malu Zhang, Shuai Wang, Jingya Wang, Wenjie Wei, Zeyu Ma, Guoqing Wang, Yang Yang, and Haizhou Li. Dendritic resonate-and-fire neuron for effective and efficient long sequence modeling. In *NeurIPS*, pages 1–27, 2025.