Fed-AGA: An Anchor Graph Alignment Framework for Federated Unaligned Multi-view Clustering

Yichi Zhang[1], Bohang Sun[2], Hao Wei[3], Kai Di[4], Zhen Yang[3], Gengyu Lyu[3]*

[1]School of Mathematics, Statistics and Mechanics, Beijing University of Technology
[2]School of Computer Science and Technology, Beijing Institute of Technology
[3]College of Computer Science, Beijing University of Technology
[4]Hangzhou School of Automation, Zhejiang Normal University

yczhang@emails.bjut.edu.cn, sunbohang2000@163.com, haowei@emails.bjut.edu.cn
dikai@seu.edu.cn, yangzhen@bjut.edu.cn, lyugengyu@gmail.com

Abstract

Federated Multi-view Clustering (FMVC) enables privacy-preserving cross-view fusion from distributed multi-view data, where each client holds one view of samples and participates in the server's fusion without exposing raw data. In practice, existing FMVC methods face two key challenges: (1) the collected samples often show unaligned correspondence across views, especially under distributed collection; and (2) the gap between higher data privacy and lower computation cost remains hard to reconcile. To address these issues, we propose an Anchor Graph Alignment Framework for Federated Unaligned Multi-view Clustering (**Fed-AGA**), which employs a coarse-to-fine alignment strategy on lightweight anchor graphs to achieve misalignment-robust cross-view fusion while bridging the gap between data privacy and computation cost. Specifically, in the coarse alignment stage, we introduce an anchor-based graph generation module to generate an anchor graph on each client and upload them to the server for category-wise alignment by a designed topology-aware alignment module, which distills topology components as category-exclusive signatures to guide and achieve ideal category-wise alignment. In the next fine alignment and fusion stage, we design an attention-based alignment module to encourage each sample to pay attention to highly similar samples for sample-wise alignment, and the fine-aligned results are contrastively fused into a cross-view consistent anchor graph for clustering. Extensive experiments on multiple datasets demonstrate that Fed-AGA achieves state-of-the-art performance among related methods, while theoretical analysis offers theoretical motivation for two key components.

1 Introduction

Multi-view data are often acquired from multiple sources and described by several heterogeneous features. For instance, in cross-platform personalized recommendation, users' activity histories are often distributed across multiple platforms. To facilitate targeted recommendation, we need to fully leverage the complementary information from these sources. However, in practical scenarios, multi-view data are often collected independently, while strict privacy regulations prohibit centralizing raw data for joint analysis. Under such conditions, traditional centralized multi-view learning methods lose their capability to effectively conduct cross-source information fusion, because they require directly accessing and aggregating raw multi-view data Fu et al. [2022], Wang et al. [2023, 2024] .

*Gengyu Lyu is the corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The key to learning from such privacy-sensitive multi-view data lies in how to fuse these heterogeneous cross-source features Lyu et al. [2022, 2024] efficiently while preserving the data privacy of each independent data holder. Recently, federated multi-view learning provides an effective solution, which deploys multi-view clustering into the federated learning framework Fu et al. [2025a,b] and enables the server to perform cross-view fusion by aggregating uploaded latent representations, without accessing the raw multi-view data. For instance, Jiang et al. [2024] proposes a feature-based FMVC method, which generates latent representations on each client using a deep autoencoder and then coordinates sample-level cross-view fusion through a server, using one-to-one cross-view counterparts to pull paired samples closer. In this process, sharing such feature-level information is vulnerable to model inversion attacks Sun et al. [2021], leading to sensitive information leakage. To mitigate these privacy risks, Sun et al. [2025] alternatively proposes a graph-based FMVC method that shares view-specific diversity graphs encoding only sample-wise correlations, while excluding latent feature information, thereby reducing the risk of privacy leakage. The method constructs a view-specific diversity graph on each client and uploads it to the server, where the server directly fuses these local structures to obtain a unified global graph representation for sample-level cross-view fusion. However, the above FMVC methods suffer from two common drawbacks: (1) they require strict one-to-one cross-view sample alignment for direct fusion, whereas real-world data are often unaligned especially under distributed collection; and (2) they face a privacy–efficiency dilemma: exchanging latent features weakens privacy protection, whereas sharing dense graphs increases computation cost.

To address these challenges, we propose a novel Anchor Graph Alignment Framework for Federated Unaligned Multi-view Clustering (Fed-AGA), which employs a coarse-to-fine alignment strategy on lightweight anchor graphs to enable misalignment-robust fusion while mitigating privacy leakage risk and reducing computation cost. Specifically, in the coarse alignment stage, each client constructs a lightweight anchor graph and uploads it to the server for category-wise alignment. On the server, we design a theoretically justified reference-view selection strategy using silhouette scores to choose a clearer alignment reference for cross-view alignment. Then, we design a topology-aware alignment module to distill cluster-level topology as category-exclusive signatures, recover category correspondences via cross-view signature similarity, and guide clients to refine anchor graphs. In the fine alignment and fusion stage, we design an attention-based alignment module with cluster priors from the coarse alignment stage, which incorporates them as a soft bias into cross-view attention to recover sample-wise correspondences, and contrastively fuses the aligned anchor graphs into a consensus anchor graph for clustering. In summary, our main contributions lie in the following aspects:

- We propose a novel Anchor Graph Alignment Framework for Federated Unaligned Multi-view Clustering, which is the first attempt to perform cross-view fusion in the **fully unaligned** FMVC scenario while maintaining higher privacy and lower computation cost.
- We mitigate the gap between higher privacy preservation and lower computation cost by utilizing lightweight anchor graphs. On these graphs, we design reference-view selection and topology-aware alignment strategies to build category-correspondence priors, which inform our attention-based alignment module, facilitating sample-wise alignment and fusion.
- We conduct extensive experiments on multiple benchmarks to demonstrate Fed-AGA's state-of-the-art performance and specifically validate its effectiveness in mitigating the gap between higher privacy preservation and lower computation cost. We also provide theoretical analyses to support our reference-view selection and topology-aware alignment strategies.

2 Related Work

2.1 Federated Multi-View Clustering

Federated multi-view clustering (FMVC) aims to perform global clustering on multi-view data distributed across multiple clients by fusing cross-view information under privacy constraints Huang et al. [2022], Chen and Zhang [2022], Ren et al. [2024], Hu et al. [2024]. For instance, Chen et al. [2023] propose a deep FMVC approach that utilizes deep autoencoders on clients to learn view-specific clustering representations, based on which the server constructs global prototypes and assigns pseudo-labels for clustering. Sun et al. [2025] develop a self-representation based framework in which each client constructs a view-specific diversity graph via self-representation and sends it to the server for cross-view structure fusion. Although these methods have achieved significant advances

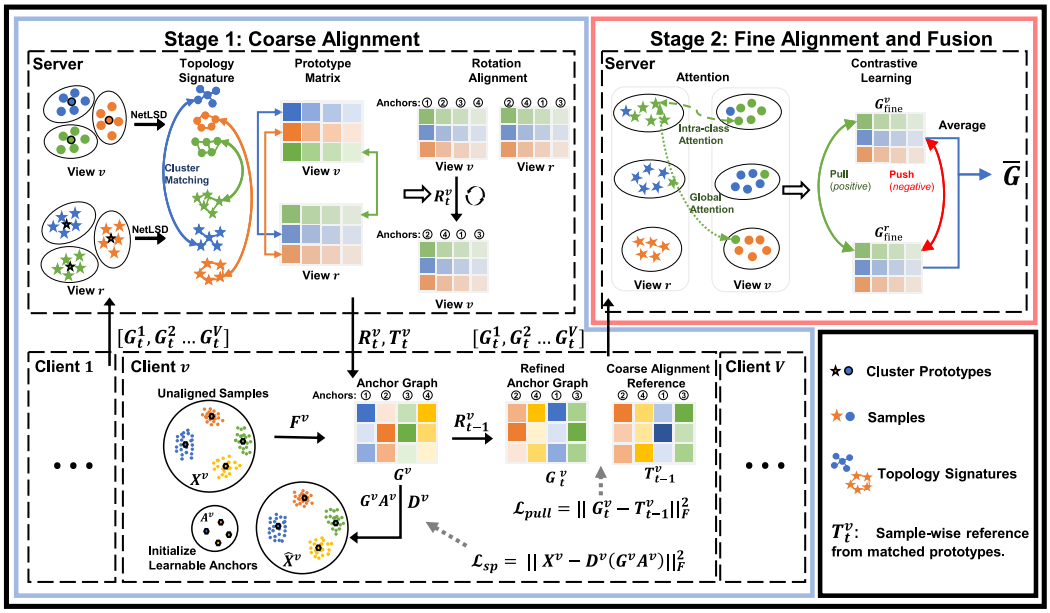


Figure 1: The overview of Fed-AGA, which follows a progressive coarse-to-fine alignment paradigm consisting of two stages. In **Stage 1 (Coarse Alignment)**, at communication round t, each client receives the previous round's coarse alignment reference $\mathbf{T}_{t-1}^v$ to guide its local training and generate its refined anchor graph $\mathbf{G}_t^v$, which is also rotated by the rotation matrix $\mathbf{R}_{t-1}^v$ and then uploaded to the server. Based on the collected $\{\mathbf{G}_t^v\}_{v=1}^V$, the server infers cross-view category correspondences and produces updated $\mathbf{R}_t^v$ and $\mathbf{T}_t^v$ for the next round training. After several rounds of coarse alignment, $\{\mathbf{G}_t^v\}_{v=1}^V$ are treated as coarsely aligned anchor graphs and sent to the next stage. **Stage 2 (Fine Alignment and Fusion)** integrates into the subsequent communication rounds. The server further conducts cross-view attention to recover sample-wise correspondences on the uploaded coarsely aligned anchor graphs and produce fine-aligned graphs $\{\mathbf{G}_{\text{fine}}^v\}_{v=1}^V$, which are further contrastively fused into a cross-view consistent consensus anchor graph $\bar{\mathbf{G}}$ for clustering.

in FMVC, they are still trapped in a privacy–efficiency dilemma: exchanging latent features weakens privacy protection, whereas sharing dense graphs increases computation cost Lu et al. [2023].

2.2 Unaligned Multi-View Clustering

Unaligned multi-view clustering (UMC) studies multi-view clustering in a fully unaligned scenario, where samples are mismatched across views Lin et al. [2022], Xin et al. [2023], Lyu et al. [2025], Zhong et al. [2025]. Existing UMC methods first estimate cross-view alignment and then leverage it for fusion. They can be grouped into structure-based and feature-based paradigms. For instance, Ji et al. [2024] propose a feature-based method that learns multi-scale representation matrices and establishes cross-view mapping matrices in multi-scale spaces via spectral relaxation, followed by a tensorial multi-scale fusion module to capture high-order cross-view correlations. Jiang et al. [2026] propose a structure-based approach that constructs view-specific bipartite anchor graphs and introduces view-specific alignment matrices to rearrange the shuffled sample order; then anchors are merged to construct a joint bipartite graph for efficient clustering. Although these methods have advanced significantly, they all operate in a centralized manner and require joint access to multi-view data or learned representations, making them unsuitable for privacy-preserving federated scenarios.

3 The Proposed Method

In this paper, we propose **Fed-AGA**, an anchor graph alignment framework for federated unaligned multi-view clustering. The core idea of Fed-AGA is to formulate unaligned cross-view fusion as a coarse-to-fine alignment process: it first recovers category correspondences from stable cluster-level structures and then uses them to facilitate sample-level alignment and fusion. Accordingly, Fed-AGA

adopts **lightweight anchor graphs** for client-server interaction, reducing the privacy leakage risk of latent-feature transmission and the computational cost of dense sample-wise graphs. The overall framework consists of two stages: **coarse alignment** and **fine alignment and fusion**. In the coarse alignment stage, we design a **topology-aware alignment module**, which distills stable cluster-level topology on anchor graphs into category-exclusive structural features to recover cross-view category correspondences. In the fine alignment and fusion stage, we design an **attention-based alignment module** that incorporates these correspondences into cross-view attention as cluster-level priors, encouraging the model to focus more on reliable cross-view samples and facilitating sample-level alignment and fusion. Finally, the fine-aligned anchor graphs are contrastively fused into a consensus anchor graph $\bar{\mathbf{G}}$ for final clustering. Figure 1 illustrates the framework of Fed-AGA.

3.1 Problem Definition

In federated unaligned multi-view clustering, multi-view data are distributed across multiple clients, and a central server is employed to coordinate them to achieve privacy-preserving cross-view clustering without sharing raw data. Formally, multi-view data with V views are denoted as $\mathbf{X} = \{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^V\}$ and distributed across V clients. For each client v, its data are represented as $\mathbf{X}^v = \{\mathbf{x}_i^v\}_{i=1}^N \in \mathbb{R}^{N \times D_v}$, where N is the number of samples and D_v is the feature dimensionality of view v, $v = 1, \ldots, V$. The goal of federated unaligned multi-view clustering is to achieve desired cross-view clustering by fusing multi-view information in samples $\{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^V\}$. Notably, the multi-view sample features cannot be directly shared and fused since: (1) the data privacy in each client is strictly emphasized and preserved, (2) $\{(\mathbf{x}_i^p, \mathbf{x}_i^q), p \neq q\}$ are often derived from different samples in the unaligned multi-view dataset.

3.2 Coarse Alignment

During coarse alignment, at communication round t, clients refine local anchor graphs $\mathbf{G}_t^v$ based on the last round guidance $\{\mathbf{T}_{t-1}^v,\ \mathbf{R}_{t-1}^v\}$ and upload $\mathbf{G}_t^v$ to the server. The server then performs coarse alignment and generates updated guidance $\{\mathbf{T}_t^v,\ \mathbf{R}_t^v\}$ for the next-round training.

Client Training. In each client v, we perform local training to generate a refined anchor graph $\mathbf{G}_t^v$, which is then uploaded to the server for subsequent coarse alignment. Specifically, we first randomly initialize a set of learnable anchors, represented by an anchor matrix $\mathbf{A}^v \in \mathbb{R}^{M \times d}$, where M denotes the number of anchors and d denotes the anchor dimension. We then design an anchor graph generation module $\mathbf{F}^v(\cdot)$ to model the sample-to-anchor relations and obtain the corresponding anchor graph $\mathbf{G}^v \in \mathbb{R}^{N \times M}$, where $\mathbf{G}^v = \mathbf{F}^v(\mathbf{X}^v)$, and $\mathbf{F}^v(\cdot)$ is implemented as a neural network. To further improve the quality of $\mathbf{G}^v$, we design a structure reconstruction module $\mathbf{D}^v(\cdot)$, which reconstructs $\mathbf{X}^v$ from $\mathbf{G}^v\mathbf{A}^v$, encouraging $\mathbf{G}^v$ to better characterize the sample–anchor relations, where $\hat{\mathbf{X}}^v = \mathbf{D}^v(\mathbf{G}^v\mathbf{A}^v)$, and $\mathbf{D}^v(\cdot)$ is also implemented as a neural network. The corresponding training loss is formulated as:

$$\mathcal{L}_{\mathrm{sp}}^v = \left\| \mathbf{X}^v - \hat{\mathbf{X}}^v \right\|_F^2 = \left\| \mathbf{X}^v - \mathbf{D}^v(\mathbf{G}^v\mathbf{A}^v) \right\|_F^2, \tag{1}$$

where the anchor graph $\mathbf{G}^v \in \mathbb{R}^{N \times M}$ with $M \ll N$ is lighter than a commonly used sample-to-sample affinity matrix in $\mathbb{R}^{N \times N}$, substantially reducing computation cost in federated scenarios.

To provide the server with higher-quality inputs for effective coarse alignment, we refine the local anchor graphs $\mathbf{G}^v$ across communication rounds. At each communication round $t (t > 1)$, each client receives the coarse alignment guidance $\{\mathbf{T}_{t-1}^v,\ \mathbf{R}_{t-1}^v\}$ computed by the server in the previous round $t-1$, where $\mathbf{T}_{t-1}^v \in \mathbb{R}^{N \times M}$ serves as the coarse alignment reference for guiding each client to refine its local anchor graph, thereby providing high-quality inputs for more effective server-side coarse alignment. The pull loss is formulated as:

$$\mathcal{L}_{\mathrm{pull}}^v = \left\| \mathbf{G}^v\mathbf{R}_{t-1}^v - \mathbf{T}_{t-1}^v \right\|_F^2, \tag{2}$$

where $\mathbf{R}_{t-1}^v \in \mathbb{R}^{M \times M}$ is the server-computed orthogonal column-wise transformation for facilitating the pulling of $\mathbf{G}^v$ toward $\mathbf{T}_{t-1}^v$, and $\mathbf{G}_t^v = \mathbf{G}^v\mathbf{R}_{t-1}^v$ is uploaded as the refined anchor graph. The client-side coarse alignment loss is then defined as:

$$\mathcal{L}_{\mathrm{coarse}}^v = \mathcal{L}_{\mathrm{sp}}^v + \lambda\, \mathcal{L}_{\mathrm{pull}}^v, \tag{3}$$

where λ is a trade-off parameter, and $\mathcal{L}_{\text{pull}}^v$ is omitted when $t = 1$.

Server Training. In the server, we receive the refined anchor graphs $\{\mathbf{G}_t^v\}_{v=1}^V$ from clients and perform cluster-level coarse alignment. Specifically, we select the view with the highest silhouette coefficient (SC) score Rousseeuw [1987] as the reference view r and align all other views to r to establish cross-view correspondences. A higher silhouette coefficient score indicates more compact clusters and better inter-cluster separation, and we further establish an SC-clarity bound in Theorem 1 that theoretically supports this criterion for selecting a reliable reference view for coarse alignment.

Theorem 1. *The SC score provides a lower bound on a view's clustering clarity: higher values indicate more compact and better-separated clusters, reducing cross-cluster confusion.*

The detailed proof is provided in Appendix A. Reference-view analysis is in Appendix C.

With the high-quality reference view r selected, we perform cluster matching between each view v and r to obtain cross-view correspondences. To this end, we cluster-wise extract the topological structural components as cross-view comparable topology signatures. We first apply k-means on each refined anchor graph $\mathbf{G}_t^v$ to obtain a view-specific partition $\mathcal{C}^v = \{\mathcal{C}_1^v, \mathcal{C}_2^v, \ldots, \mathcal{C}_K^v\}$, where $\mathcal{C}_k^v$ denotes the k-th cluster in view v. To enhance the reliability of these partitions, we retain only the top $p\%$ samples closest to the prototype μ_k^v within each $\mathcal{C}_k^v$. These highly reliable samples form $\mathbf{B}_k^v \in \mathbb{R}^{n_k^v \times M}$ for subsequent signature construction, where n_k^v is the number of retained samples.

To extract the topology signatures, we first compute the pairwise similarity matrix $\mathbf{W}_k^v = \mathbf{B}_k^v(\mathbf{B}_k^v)^\top \in \mathbb{R}^{n_k^v \times n_k^v}$, which is then symmetrized as $\widetilde{\mathbf{W}}_k^v = \frac{1}{2}\big(\mathbf{W}_k^v + (\mathbf{W}_k^v)^\top\big)$, and we compute the normalized Laplacian as $\mathbf{L}_k^v = \mathbf{I} - (\mathbf{D}_k^v)^{-1/2}\widetilde{\mathbf{W}}_k^v(\mathbf{D}_k^v)^{-1/2}$, where $(\mathbf{D}_k^v)_{ii} = \sum_j (\widetilde{\mathbf{W}}_k^v)_{ij}$. Finally, given the normalized Laplacian $\mathbf{L}_k^v$, we adopt the *Network Laplacian Spectral Descriptor* (NetLSD) Tsitsulin et al. [2018] to extract a stable and cross-view comparable topology signature $\mathbf{s}_k^v$ from $\mathbf{L}_k^v$ for our topology-guided cluster matching, as we theoretically establish in Theorem 2.

Theorem 2. *The NetLSD topology signature is permutation-invariant and stable under small cluster-level spectral perturbations, thereby supporting robust cluster comparison for coarse alignment.*

The detailed proof is provided in Appendix B.

Given these signatures, for cluster alignment, we evaluate the pairwise cluster dissimilarity to construct a cost matrix $\mathbf{C}^{rv} \in \mathbb{R}^{K\times K}$, where $\mathbf{C}_{k\ell}^{rv} = \|\mathbf{s}_k^v - \mathbf{s}_\ell^r\|_2$ and larger values correspond to more dissimilar clusters. We then minimize the total cost $\operatorname{tr}\big(\mathbf{P}(\mathbf{C}^{rv})^\top\big)$ via the Hungarian algorithm to match highly similar clusters, yielding the permutation matrix $\mathbf{P}^{rv}$:

$$\mathbf{P}^{rv} = \arg\min_{\mathbf{P}\in\{0,1\}^{K\times K}} \operatorname{tr}\big(\mathbf{P}(\mathbf{C}^{rv})^\top\big) \quad \text{s.t. } \mathbf{P}\mathbf{1} = \mathbf{1},\ \mathbf{P}^\top\mathbf{1} = \mathbf{1}. \tag{4}$$

where $\mathbf{P}_{k\ell}^{rv} = 1$ means that the ℓ-th cluster in the reference view r is matched to the k-th cluster in view v, completing the cluster-level coarse alignment. To leverage server-side coarse alignment information for refining local anchor graphs, we generate guidance $\{\mathbf{T}_t^v,\ \mathbf{R}_t^v\}$ for client-side anchor graph optimization. Specifically, we obtain the prototype matrix $\mathbf{H}^v = [\boldsymbol{\mu}_1^v; \ldots; \boldsymbol{\mu}_K^v] \in \mathbb{R}^{K\times M}$ and the indicator matrix $\mathbf{Y}^v \in \{0,1\}^{N\times K}$ from the partition of $\mathbf{G}_t^v$, where $\mathbf{Y}_{ik}^v = 1$ indicates that sample i is assigned to cluster k. Using the permutation matrix $\mathbf{P}^{rv}$, we re-order the reference prototypes as $\tilde{\mathbf{H}}^v = \mathbf{P}^{rv}\mathbf{H}^r \in \mathbb{R}^{K\times M}$. We then compute $\{\mathbf{T}_t^v,\ \mathbf{R}_t^v\}$ as follows.

(1) Generation of the coarse alignment reference $\mathbf{T}_t^v$: To provide an alignment reference and guide the local anchor graph $\mathbf{G}_t^v$ toward the corresponding prototypes, we construct $\mathbf{T}_t^v$ as:

$$\mathbf{T}_t^v = \mathbf{Y}^v\mathbf{P}^{rv}\mathbf{H}^r \in \mathbb{R}^{N\times M}, \tag{5}$$

where each row of $\mathbf{T}_t^v$ provides the matched reference-view prototype for the corresponding sample.

(2) Generation of the rotation matrix $\mathbf{R}_t^v$: To further facilitate pulling the anchor graph $\mathbf{G}_t^v$ toward $\mathbf{T}_t^v$, we compute $\mathbf{R}_t^v$ as an orthogonal transformation to mitigate cross-view column discrepancies:

$$\mathbf{R}_t^v = \arg\min_{\mathbf{R}^\top\mathbf{R}=\mathbf{I}} \left\|\mathbf{H}^v\mathbf{R} - \tilde{\mathbf{H}}^v\right\|_F^2, \tag{6}$$

where $\mathbf{H}^v$ and $\tilde{\mathbf{H}}^v$ are regarded as the core components of $\mathbf{G}_t^v$ and $\mathbf{T}_t^v$, respectively. As $\mathbf{R}_t^v$ is optimized to pull $\mathbf{H}^v$ toward $\tilde{\mathbf{H}}^v$, this transformation can likewise be applied to the entire anchor graph $\mathbf{G}_t^v$, thereby facilitating the pulling of $\mathbf{G}_t^v$ toward the alignment reference $\mathbf{T}_t^v$.

The server then distributes $\{\mathbf{T}_t^v,\ \mathbf{R}_t^v\}$ to each non-reference client for refining local anchor graphs in the next round. After several rounds of coarse alignment, the refined anchor graphs $\mathbf{G}_t^v$ are treated as coarsely aligned anchor graphs and sent to the fine alignment and fusion stage.

3.3 Fine Alignment and Fusion

After receiving the refined coarsely aligned anchor graphs $\{\mathbf{G}_t^v\}_{v=1}^V$, we design an attention-based alignment module that incorporates the recovered cluster-level correspondences as soft priors into cross-view attention, encouraging each sample to focus on more reliable cross-view candidates. The fine-aligned representations are then contrastively fused into a consensus anchor graph for clustering.

Attention-based alignment. We employ an attention mechanism to achieve fine-grained sample-wise alignment. Specifically, we compute the query feature $\mathbf{Q}$ for the reference view and the key–value features $\mathbf{K}^v, \mathbf{V}^v$ for each non-reference view v, where $\mathbf{Q} = \phi(\mathbf{G}_t^r)$ and $\mathbf{K}^v = \mathbf{V}^v = \psi^v(\mathbf{G}_t^v)$, with $\phi(\cdot)$ and $\psi^v(\cdot)$ denoting projection networks. To reduce redundancy, a tied key–value design is adopted for $\psi^v(\cdot)$, as commonly used in attention designs Lee et al. [2019]. We then compute the attention matrix $\mathbf{U}^v$ to obtain cross-view similarity weights between $\mathbf{Q}$ and $\mathbf{K}^v$, enabling the reference-view queries to focus on reliable samples in other views, thereby facilitating alignment:

$$\mathbf{U}^v = \frac{\mathbf{Q}(\mathbf{K}^v)^\top}{\sqrt{d_k}}. \tag{7}$$

To achieve robust fine-grained alignment and suppress irrelevant correspondences, we refine the attention scores by integrating a cluster-level bias with a top-k selection strategy. Specifically, to encourage each sample to focus more on samples from the same category, we construct an indicator matrix $\mathbf{\Omega}^v \in \{0,1\}^{N\times N}$ to define where the attention bias α should be applied. Given the coarse-alignment permutation matrix $\mathbf{P}^{rv}$, we set $\Omega_{ij}^v = 1$ when reference sample i and view-v sample j are assigned to matched clusters, and 0 otherwise. Subsequently, as standard attention assigns weights to all samples, including unrelated ones, we adopt a top-k mechanism Gupta et al. [2021] to filter out these erroneous correspondences and prevent incorrect alignment. The overall calculation process for the refined attention scores is formulated as:

$$\tilde{\mathbf{U}}^v = \mathbf{TopK}\left(\frac{\mathbf{Q}(\mathbf{K}^v)^\top}{\sqrt{d_k}} + \alpha\,\mathbf{\Omega}^v\right), \tag{8}$$

where d_k denotes the projection dimension and $\alpha > 0$ controls the strength of the cluster-level guidance. Also, the $\mathbf{TopK}(\cdot)$ operator performs row-wise sparsification, retaining only the k largest similarity scores while setting the rest to $-\infty$. Then we apply a row-wise softmax on $\tilde{\mathbf{U}}^v$ to obtain the attention matrix $\mathbf{S}^v$. The fine-aligned anchor graph of view v is then obtained as: $\mathbf{G}_{\text{fine}}^v = \omega(\mathbf{S}^v\mathbf{V}^v)$, where $\omega(\cdot)$ denotes the output projection.

Global contrastive objective and fusion. After we obtain a set of fine-aligned anchor graphs $\{\mathbf{G}_{\text{fine}}^v\}_{v=1}^V$, we then perform cross-view contrastive fusion to enhance sample-wise consistency. For each view v, $\mathbf{g}_{\text{fine},i}^v$ is the i-th row of $\mathbf{G}_{\text{fine}}^v$. And for each non-reference view $v \neq r$, we construct contrastive pairs with the reference view r: the positive pair is defined as $(\mathbf{g}_{\text{fine},i}^r, \mathbf{g}_{\text{fine},i}^v)$, whereas the negative pair is $(\mathbf{g}_{\text{fine},i}^r, \mathbf{g}_{\text{fine},j}^v)$ for all $j \neq i$. We then introduce a contrastive loss to enhance the consistency of positive pairs while separating mismatched negatives, enabling cross-view fusion. The whole loss function is represented as:

$$\mathcal{L}_{\text{fine}} = \sum_{v\neq r}\frac{1}{N}\sum_{i=1}^{N} -\log\frac{e^{\text{sim}(\mathbf{g}_{\text{fine},i}^r, \mathbf{g}_{\text{fine},i}^v)/\tau}}{\sum_{j=1}^N e^{\text{sim}(\mathbf{g}_{\text{fine},i}^r, \mathbf{g}_{\text{fine},j}^v)/\tau}}, \tag{9}$$

where τ is a hyperparameter and $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity. This loss enforces sample-wise consistency on the fine-aligned anchor graphs, thereby strengthening cross-view correspondences and facilitating cross-view fusion.

Finally, we obtain the fine-aligned anchor representations $\{\mathbf{G}_{\text{fine}}^v\}_{v=1}^V$ from all views. These representations are averaged to obtain the consensus anchor graph $\bar{\mathbf{G}} = \frac{1}{V}\sum_{v=1}^V \mathbf{G}_{\text{fine}}^v$, on which we perform k-means to obtain the final clustering results.

Table 1: Statistical characteristics of the seven datasets.

Data	Samples	Clusters	View dimensions
Caltech7-5	1400	7	40/254/1984/512/928
Scene	4485	15	20/59/40
Caltech102-5	9144	102	48/40/254/512/928
Hdigit	10000	10	784/256
Aloi	10800	100	77/13/64/125
NUSWIDEOBJ	30000	31	65/226/145/74/129
Cifar10	50000	10	512/2048/1024

3.4 Optimization

In each communication round t of the coarse alignment stage, clients refine their anchor graphs under the server-provided $\{\mathbf{T}_{t-1}^v, \mathbf{R}_{t-1}^v\}$ by minimizing Eq. (3) and upload the refined anchor graph $\mathbf{G}_t^v$ to the server for coarse alignment. After the server receives $\{\mathbf{G}_t^v\}_{v=1}^V$, it selects the reference view r at the first round and keeps it fixed thereafter. The server then performs coarse alignment and produces coarse alignment guidance $\{\mathbf{T}_t^v, \mathbf{R}_t^v\}$ for each client $v \neq r$ to guide the next round of client-side refinement. After several coarse alignment rounds, we integrate the fine alignment and fusion stage into the subsequent communication rounds. In this stage, the server employs an attention-based alignment strategy by optimizing Eq. (9) to obtain fine-aligned graphs, which are then fused into a consensus anchor graph $\bar{\mathbf{G}}$ for k-means clustering to obtain the final results.

4 Experiments

In this section, we evaluate Fed-AGA on clustering performance, component effectiveness, convergence, sensitivity, computation cost, and privacy leakage risk. Appendix C reports auxiliary studies, such as reference-view selection, extra ablations, and extended cost–privacy results.

4.1 Experimental Settings

Datasets. We employ seven widely-used multi-view datasets for comparative studies, including ***Caltech7-5*** Fei-Fei et al. [2004], ***Scene*** Fei-Fei and Perona [2005], ***Caltech102-5*** Fei-Fei et al. [2004], ***Hdigit*** Chen et al. [2022], ***Aloi*** Li et al. [2023], ***NUSWIDEOBJ*** Li et al. [2015], and ***Cifar10*** Zhang et al. [2018]. The specific characteristics of these datasets are recorded in Table 1.

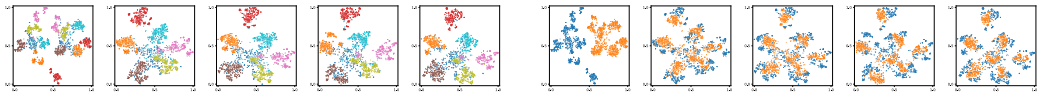
The Compared Methods. To verify the effectiveness of our proposed method, we compare it with seven state-of-the-art methods, including five unaligned multi-view clustering methods: ***MVC-UM*** Yu et al. [2021], ***UPMGC*** Wen et al. [2023], ***OpVuC*** Dong et al. [2024], ***TUMCR*** Ji et al. [2024], ***MGCCFF*** Zhao et al. [2025] and two federated methods: ***FedDMVC*** Chen et al. [2023] and ***MGCD*** Sun et al. [2025], where all compared methods are implemented according to the source codes released by the authors, and the optimal parameters are set following the original papers.

Metrics. There are four widely-used metrics applied to quantitatively evaluate the performance of multi-view clustering methods, including Accuracy (ACC), Normalized Mutual Information (NMI), Purity (PUR) and Adjusted Rand Index (ARI), whose detailed definitions are illustrated in Liang et al. [2022]. For each of the above metrics, a higher value indicates better performance.

Implementation Details. The proposed method is implemented using the PyTorch Paszke et al. [2019] framework. On the client side, $F(\cdot)$ is implemented as a multi-layer fully connected network with layer dimensions $\{D, 500, 500, 2000, M\}$, where M is twice the number of clusters. And $D(\cdot)$ adopts a symmetric architecture with dimensions $\{d, 2000, 500, 500, D\}$, where d is set to 20. ReLU is used as the activation function in all layers. On the server side, the query projection is implemented as a linear layer, while the key/value and output projection networks are realized by MLPs with dimensions $\{M, 2000, M\}$. The model is optimized using Adam Kingma and Ba [2014]. The pre-training stage lasts 100 epochs, and the collaborative training proceeds for 10 global communication rounds with a mini-batch size of 256. All experiments are conducted on a workstation with an Intel(R) Xeon(R) Gold 6148 2.40GHz CPU, GeForce RTX 3090 GPUs, and 128GB RAM.

Table 2: Comparative results between our proposed Fed-AGA and 7 state-of-the-art methods on seven datasets, where the best results are presented in **bold** and the second-best results are underlined. "-" means that the code cannot be run due to excessive time or space complexity.

Dataset	Metric	UMVC					FMVC		Ours
		MVC-UM	OpVuC	TUMCR	UPMGC	MGCCFF	MGCD	FedDMVC	Fed-AGA
Scene	ACC	0.2608	0.3275	0.2990	0.1386	0.2713	0.1870	0.1701	**0.3946**
	NMI	0.2787	0.3409	0.2488	0.0576	0.3793	0.1450	0.1922	**0.3993**
	ARI	0.1947	0.1836	0.2171	0.0434	0.1092	0.0585	0.0718	**0.2206**
	PUR	0.2791	0.3741	0.3398	0.1503	0.3197	0.2031	0.2161	**0.4430**
Caltech7-5	ACC	0.2785	0.3279	0.2557	0.8079	0.3857	0.2594	0.3421	**0.8807**
	NMI	0.1038	0.1229	0.0721	0.7137	0.2543	0.1057	0.2146	**0.7902**
	ARI	0.0875	0.0820	0.0612	0.7034	0.2002	0.0632	0.1805	**0.7683**
	PUR	0.3014	0.3851	0.2771	0.8079	0.4381	0.2656	0.3450	**0.8807**
Caltech102-5	ACC	0.0638	0.1355	0.0977	0.0930	0.1708	0.0429	0.0551	**0.2266**
	NMI	0.2150	0.2974	0.2066	0.1883	0.4085	0.1410	0.1814	**0.4421**
	ARI	0.0571	0.0958	0.0639	0.0742	0.0335	0.0013	0.0112	**0.1810**
	PUR	0.1846	0.2794	0.1943	0.1704	0.3462	0.1161	0.1580	**0.4185**
Hdigit	ACC	0.4627	0.3994	0.1551	0.4087	0.4167	0.1801	0.4099	**0.5883**
	NMI	0.4418	0.3151	0.0255	0.3700	0.4690	0.0473	0.3421	**0.5224**
	ARI	0.3987	0.2054	0.0203	0.2943	0.0626	0.0314	0.2463	**0.4168**
	PUR	0.5031	0.4173	0.1679	0.4545	0.5182	0.1831	0.4231	**0.6303**
Aloi	ACC	0.3543	0.1432	0.2330	0.0383	0.4176	0.0987	0.0567	**0.4878**
	NMI	0.6039	0.4105	0.3363	0.1118	0.6877	0.2636	0.1824	**0.6950**
	ARI	0.2381	0.0897	0.1853	0.0213	0.0747	0.0371	0.0102	**0.3590**
	PUR	0.3692	0.1747	0.2429	0.0402	0.4252	0.1154	0.0599	**0.5216**
NUSWIDEOBJ	ACC	-	0.1016	-	-	-	0.0449	0.0585	**0.1129**
	NMI	-	0.0864	-	-	-	0.0063	0.0248	**0.0923**
	ARI	-	0.0213	-	-	-	0.0002	0.0047	**0.0252**
	PUR	-	**0.2041**	-	-	-	0.1225	0.1416	0.2040
Cifar10	ACC	-	0.8008	-	-	-	0.4623	0.6128	**0.8809**
	NMI	-	0.6872	-	-	-	0.3582	0.5687	**0.7674**
	ARI	-	0.6284	-	-	-	0.2662	0.4904	**0.7538**
	PUR	-	0.8008	-	-	-	0.4660	0.6148	**0.8809**



(a) Colored by cluster, rounds 0, 2, 4, 6, 8 (b) Colored by view, rounds 0, 2, 4, 6, 8

Figure 2: Evolution of coarsely aligned anchor graphs $\mathbf{G}_t$ on Caltech7-5 at communication rounds 0, 2, 4, 6, and 8. (a) Points are colored by cluster labels; closer same-cluster samples across views indicate clearer cross-view cluster correspondence. (b) Points are colored by view labels; closer and more thoroughly mixed representations from different views indicate stronger cross-view consistency.

4.2 Experimental Results

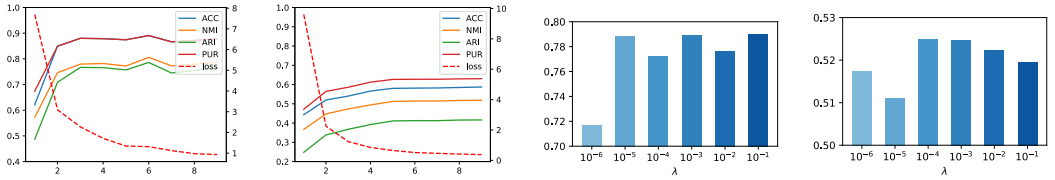
Comparisons with other methods. Table 2 records the experimental comparisons between our proposed Fed-AGA method and 7 state-of-the-art methods, where the best results are highlighted in **bold**, and the second-best are underlined. In addition, Figure 2 illustrates the visualization of coarsely aligned results of our method on Caltech7-5. According to Table 2 and Figure 2, we can observe that:

(1) Among all the employed datasets, our Fed-AGA is superior to almost all competing methods on all evaluation metrics, and even achieves a significant lead over the sub-optimal methods. In particular, on Hdigit, Fed-AGA improves over the second-best method by 12.56%, 5.34%, 1.81% and 11.21% on ACC, NMI, ARI and PUR, respectively. These comparative results verify the effectiveness of Fed-AGA for federated clustering under fully unaligned views.

(2) Figure 2 visualizes the evolution of coarsely aligned anchor graphs $\mathbf{G}_t$ on Caltech7-5 every two communication rounds. When colored by cluster labels, same-cluster samples across views gradually move closer, indicating clearer cross-view cluster correspondence. When colored by view labels, representations from different views become increasingly mixed, showing stronger cross-view

Table 3: Ablation study on key components on *Caltech7-5* dataset.

	$\mathcal{L}_{sp}$	$\mathbf{R}$	$\mathcal{L}_{pull}$	$\mathcal{FA}$	ACC	NMI	ARI	PUR
(A)	✓				0.4157	0.2155	0.1693	0.4400
(B)	✓	✓	✓		0.3436	0.1093	0.0768	0.3450
(C)	✓			✓	0.8100	0.7103	0.6639	0.8100
(D)	✓		✓	✓	0.8336	0.7312	0.6759	0.8336
(E)	✓	✓		✓	0.8664	0.7586	0.7358	0.8664
(F)	✓	✓	✓	✓	**0.8807**	**0.7902**	**0.7683**	**0.8807**



(a) Caltech7-5 (b) Hdigit (c) Caltech7-5 (d) Hdigit

Figure 3: Convergence and sensitivity analysis on *Caltech7-5* and *Hdigit* datasets

consistency. These observations confirm that the proposed coarse alignment stage can discover reliable cluster correspondences and provide a better basis for fine alignment.

4.3 Model Analysis

Ablation Study. We conduct ablation experiments on Caltech7-5 to evaluate the contributions of key components in Fed-AGA. According to Table 3, (F) performs better than (B), suggesting that introducing $\mathcal{FA}$ (Fine Alignment and Fusion) substantially improves the results across all evaluation metrics, and thus the fine alignment and fusion stage is crucial. Moreover, (F) outperforms both (D) and (E), indicating that removing either $\mathcal{L}_{\text{pull}}$ or $\mathbf{R}$ leads to performance degradation, where $\mathcal{L}_{\text{pull}}$ and $\mathbf{R}$ are key components of the coarse alignment stage. In addition, (C) achieves better performance than (A), confirming that $\mathcal{FA}$ alone already provides significant benefits over the basic setting, while (F) further improves upon (C), showing that incorporating coarse alignment on top of $\mathcal{FA}$ yields additional gains and facilitates the subsequent fine alignment and fusion stage. Overall, the coarse alignment components ($\mathcal{L}_{\text{pull}}$, $\mathbf{R}$) and $\mathcal{FA}$ jointly improve clustering quality.

Convergence analysis. Figure 3 shows the convergence curves of our method on Caltech7-5 and Hdigit, where the loss and clustering metrics are reported in each subfigure. As communication rounds increase, the loss drops rapidly at the beginning and then stabilizes, while the clustering metrics increase and fluctuate within a small range, verifying the convergence of the proposed method.

Parameter Sensitivity Analysis. We conduct a sensitivity study to investigate the impact of the client-side hyperparameter λ on the clustering performance of Fed-AGA. Figure 3 reports the NMI on Caltech7-5 and Hdigit when varying λ from 10^{-6} to 10^{-1}. Overall, Fed-AGA exhibits stable performance over a relatively wide range of λ, and the results are particularly robust when $\lambda \in [10^{-4}, 10^{-1}]$. This indicates that the proposed optimization is not overly sensitive to the exact choice of λ, suggesting its good practical robustness across different real-world datasets.

Computation Cost and Privacy Leakage Analysis. As shown in Table 4, we evaluate the computation cost and empirical privacy leakage risk on Aloi and Cifar10. For cost, we report host memory for traditional baselines and additionally report training time and GPU memory for deep methods. Fed-AGA reduces host memory compared with traditional methods and achieves lower training time and GPU memory than deep baselines, while remaining stable on Cifar10, where several baselines become costly or fail. For privacy, following the decoder-based reconstruction evaluation Przewięźlikowski et al. [2024], we report reconstruction MSE, since a lower MSE indicates that more original feature information can be reconstructed and thus implies higher leakage risk; our anchor graphs yield higher MSE than FedDMVC's latent representations, suggesting lower empirical leakage risk. Overall, Fed-AGA reduces excessive computation cost while lowering empirical privacy leakage risk.

Table 4: Computation cost and privacy leakage analysis on Aloi and Cifar10.

Dataset	Metric	Traditional			Deep			Priv. MSE↑	
		MVC-UM	TUMCR	UPMGC	MGCD	MGCCFF	Ours	FedDMVC	Ours
Aloi	Time (s)	-	-	-	265.04	743.56	**242.42**	0.100	**0.244**
	GPU Mem. (GB)	-	-	-	1.43	18.07	**0.70**		
	Host Mem. (GB)	13.95	4.22	4.39	2.89	**1.45**	1.59		
Cifar10	Time (s)	-	-	-	3371.08	-	**2376.29**	1.969	**8.513**
	GPU Mem. (GB)	-	-	-	4.84	OOM	**1.72**		
	Host Mem. (GB)	>128.00	>128.00	>128.00	59.05	-	**2.88**		

Priv. MSE is reported as reconstruction MSE $\times 10^3$. "-" indicates an inapplicable metric or an unavailable result due to resource limitations.

5 Conclusion

In this paper, we proposed a novel Anchor-based Fusion framework (Fed-AGA) for Federated Unaligned Multi-view Clustering, which is robust to cross-view misalignment. Compared with previous methods, our proposed method conducts cross-view fusion based on lightweight anchor graphs, thereby reducing privacy leakage risk and computation cost. Meanwhile, the progressive coarse-to-fine alignment facilitates the recovery of reliable cross-view correspondences, thereby improving clustering performance. Extensive experimental results on various datasets have verified the effectiveness of our proposed method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62306020), the National Key Research and Development Program of China (No. 2023YFB3107100), the Guangxi Science and Technology Major Program (No. AA24206003) and the Inner Mongolia Autonomous Region Science and Technology Program (No. 2025KJHZ0034).

References

J. Chen and A. Zhang. FedMSplit: Correlation-adaptive federated multi-task learning across multimodal split networks. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 87–96, 2022.

Man-Sheng Chen, Jia-Qi Lin, Xiang-Long Li, Bao-Yu Liu, Chang-Dong Wang, Dong Huang, and Jian-Huang Lai. Representation learning in multi-view clustering: A literature review. *Data Science and Engineering*, 7:225–241, 2022.

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 International Conference on Multimedia*, pages 3498–3506, 2023.

Wenhua Dong, Xiao-Jun Wu, Zhenhua Feng, Sara Atito Ali Ahmed, Muhammad Awais, and Josef Kittler. One-pass view-unaligned clustering. *IEEE Transactions on Multimedia*, 26:9699–9709, 2024.

L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 524–531, 2005.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 178–186, 2004.

Lele Fu, Zhaoliang Chen, Yongyong Chen, and Shiping Wang. Unified low-rank tensor learning and spectral embedding for multi-view subspace clustering. *IEEE Transactions on Multimedia*, 25: 4972–4985, 2022.

Lele Fu, Bowen Deng, Sheng Huang, Tianchi Liao, Shirui Pan, and Chuan Chen. Less is more: Federated graph learning with alleviating topology heterogeneity from a causal perspective. In *Forty-second International Conference on Machine Learning*, 2025a.

Lele Fu, Sheng Huang, Yanyi Lai, Tianchi Liao, Chuanfu Zhang, and Chuan Chen. Beyond federated prototype learning: Learnable semantic anchors with hyperspherical contrast for domain-skewed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16648–16656, 2025b.

Ankit Gupta, Guy Dar, Shaya Goodman, David Ciprut, and Jonathan Berant. Memory-efficient transformers via top-k attention. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 39–52, 2021.

Xiangpeilin Hu, Jiandong Qin, Yongpeng Shen, Witold Pedrycz, Xinwang Liu, and Jin Liu. An efficient federated multiview fuzzy c-means clustering method. *IEEE Transactions on Fuzzy Systems*, 32:1886–1899, 2024.

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

Jintian Ji, Songhe Feng, and Yidong Li. Tensorized unaligned multi-view clustering with multi-scale representation learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 1246–1256, 2024.

Hongyu Jiang, Hong Tao, Zhangqi Jiang, and Chenping Hou. Unaligned multi-view clustering via diversified anchor graph fusion. *Pattern Recognition*, 170:111977, 2026.

Xiaorui Jiang, Zhongyi Ma, Yulin Fu, Yong Liao, and Pengyuan Zhou. Heterogeneity-aware federated deep multi-view clustering towards diverse feature representations. In *ACM International Conference on Multimedia*, pages 9184–9193, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–15, 2014.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753, 2019.

Xingfeng Li, Zhenwen Ren, Quansen Sun, and Zhi Xu. Auto-weighted tensor schatten p-norm for robust multi-view graph clustering. *Pattern Recognition*, 134:109083, 2023.

Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page 2750–2756, 2015.

Youwei Liang, Dong Huang, Chang-Dong Wang, and Philip S. Yu. Multi-view graph learning by joint modeling of consistency and inconsistency. *IEEE Transactions on Neural Networks and Learning Systems*, 35:2848–2862, 2022.

Jia-Qi Lin, Man-Sheng Chen, Chang-Dong Wang, and Haizhang Zhang. A tensor approach for uncoupled multiview clustering. *IEEE Transactions on Cybernetics*, 54:1236–1249, 2022.

Xun Lu, Songhe Feng, Gengyu Lyu, Yi Jin, and Congyan Lang. Distance-preserving embedding adaptive bipartite graph multi-view learning with application to multi-label classification. *ACM Transactions on Knowledge Discovery from Data*, 17(2):19:1–19:21, 2023.

Gengyu Lyu, Xiang Deng, Yanan Wu, and Songhe Feng. Beyond shared subspace: A view-specific fusion for multi-view multi-label learning. In *AAAI Conference on Artificial Intelligence*, pages 7647–7654, 2022.

Gengyu Lyu, Weiqi Kang, Haobo Wang, Zheng Li, Zhen Yang, and Songhe Feng. Common-individual semantic fusion for multi-view multi-label learning. In *International Joint Conference on Artificial Intelligence*, pages 4715–4723, 2024.

Gengyu Lyu, Zhen Yang, Xiang Deng, and Songhe Feng. L-VSM: Label driven view-specific fusion for multi-view multi-label classification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):6569–6583, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 1–12, 2019.

Marcin Przewięźlikowski, Marcin Osial, Bartosz Zieliński, and Marek Śmieja. A deep cut into split federated self-supervised learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 444–459, 2024.

Yazhou Ren, Xinyue Chen, Jie Xu, Jingyu Pu, Yipeng 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*, 108:102357, 2024.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Bohang Sun, Yongjian Deng, Yuena Lin, Qiuru Hai, Zhen Yang, and Gengyu Lyu. Graph consistency and diversity measurement for federated multi-view clustering. In *AAAI Conference on Artificial Intelligence*, pages 20663–20671, 2025.

Jingwei Sun, Ang Li, Binghui Wang, Huanrui Yang, Hai Li, and Yiran Chen. Soteria: Provable defense against privacy leakage in federated learning from representation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9311–9319, 2021.

Anton Tsitsulin, Davide Mottin, Panagiotis Karras, Alexander Bronstein, and Emmanuel Müller. NetLSD: Hearing the shape of a graph. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2347–2356, 2018.

Cédric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer, 2009.

Jing Wang, Songhe Feng, Gengyu Lyu, and Zhibin Gu. Triple-granularity contrastive learning for deep multi-view subspace clustering. In *ACM International Conference on Multimedia*, pages 2994–3002, 2023.

Jing Wang, Songhe Feng, Gengyu Lyu, and Jiazheng Yuan. SURER: Structure-adaptive unified graph neural network for multi-view clustering. In *AAAI Conference on Artificial Intelligence*, pages 15520–15527, 2024.

Yi Wen, Siwei Wang, Qing Liao, Weixuan Liang, Ke Liang, Xinhang Wan, and Xinwang Liu. Unpaired multi-view graph clustering with cross-view structure matching. *IEEE Transactions on Neural Networks and Learning Systems*, 35:16049–16063, 2023.

Like Xin, Wanqi Yang, Lei Wang, and Ming Yang. Selective contrastive learning for unpaired multi-view clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 36:1749–1763, 2023.

Hong Yu, Jia Tang, Guoyin Wang, and Xinbo Gao. A novel multi-view clustering method for unknown mapping relationships between cross-view samples. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 2075–2083, 2021.

Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1774–1782, 2018.

Liang Zhao, Ziyue Wang, Xiao Wang, Zhikui Chen, and Bo Xu. Incomplete and unpaired multi-view graph clustering with cross-view feature fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22789–22794, 2025.

Qiyu Zhong, Gengyu Lyu, and Zhen Yang. Align while fusion: A generalized nonaligned multiview multilabel classification method. *IEEE Transactions on Neural Networks and Learning Systems*, 36:7627–7636, 2025.

Appendix

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

- Appendix A: Proof of theorem 1.
- Appendix B: Proof of theorem 2.
- Appendix C: Additional experimental results.

A Proof of theorem 1

The silhouette coefficient (SC) score provides a lower bound on a view's clustering clarity: a higher SC score implies tighter intra-cluster structure and better inter-cluster separation, reducing cross-cluster confusion. We first introduce **Definition A.1** and **Lemma A.2**, which are the core of the theorem proof.

Definition A.1 (Silhouette Coefficient Rousseeuw [1987])**.** Consider view v with its cluster partition $\mathcal{C}^v = \{C_1^v, C_2^v, \ldots, C_K^v\}$, where K is the number of clusters and $\mathcal{S}_v = \bigcup_{k=1}^K C_k^v$ denotes the set of all samples in view v.

For a sample $i \in C_k^v$, let $a_v(i)$ denote the average distance between i and all other samples in the same cluster C_k^v, and let $b_v(i)$ denote the minimum average distance from i to samples in any other cluster C_ℓ^v with $\ell \neq k$, i.e.,

$$a_v(i) = \frac{1}{|C_k^v| - 1} \sum_{j \in C_k^v,\, j \neq i} d_v(i,j), \qquad b_v(i) = \min_{\ell \neq k} \frac{1}{|C_\ell^v|} \sum_{j \in C_\ell^v} d_v(i,j). \tag{10}$$

where $d_v(\cdot,\cdot)$ denotes a distance metric in view v, such as the Euclidean distance. The silhouette coefficient of sample i is defined as

$$s_v(i) = \frac{b_v(i) - a_v(i)}{\max\{a_v(i),\, b_v(i)\}} \in [-1, 1]. \tag{11}$$

The view-level silhouette score is given by

$$\mathrm{SC}(v) = \frac{1}{|\mathcal{S}_v|} \sum_{i \in \mathcal{S}_v} s_v(i). \tag{12}$$

Lemma A.2 (Markov Inequality)**.** Let Z be a non-negative random variable. Then for any $\epsilon > 0$, it holds that

$$\Pr(Z \geq \epsilon) \leq \frac{\mathbb{E}[Z]}{\epsilon}. \tag{13}$$

Proof: Let i be drawn uniformly at random from $\mathcal{S}_v$. Thus, $\mathrm{SC}(v) = \mathbb{E}[s_v(i)]$ under this uniform draw.

Interpretation of negative silhouettes. If $s_v(i) < 0$, then $a_v(i) > b_v(i)$, meaning that sample i is on average closer to its nearest other cluster than to its assigned cluster, and is thus inherently ambiguous or potentially misclustered.

For any threshold $\epsilon \in (0, 1)$, define the set of unclear samples as $\mathcal{B}_\epsilon = \{i :\ s_v(i) < \epsilon\}$ and the clear samples as $\mathcal{G}_\epsilon = \{i :\ s_v(i) \geq \epsilon\}$. Accordingly, the ϵ-clarity is

$$\mathrm{Clarity}_\epsilon(v) = \Pr(i \in \mathcal{G}_\epsilon) = 1 - \Pr(i \in \mathcal{B}_\epsilon) = 1 - \Pr(s_v(i) < \epsilon). \tag{14}$$

Define a nonnegative random variable $Z = 1 - s_v(i) \geq 0$. Then $s_v(i) < \epsilon \iff 1 - s_v(i) > 1 - \epsilon \iff Z > 1 - \epsilon$. Applying Lemma A.2 (Markov inequality) yields

$$\Pr(s_v(i) < \epsilon) = \Pr(Z > 1 - \epsilon) \leq \Pr(Z \geq 1 - \epsilon) \leq \frac{\mathbb{E}[Z]}{1 - \epsilon}. \tag{15}$$

Moreover, $\mathbb{E}[Z] = \mathbb{E}[1 - s_v(i)] = 1 - \mathbb{E}[s_v(i)] = 1 - \mathrm{SC}(v)$. Therefore,

$$\mathrm{Clarity}_\epsilon(v) = 1 - \Pr(s_v(i) < \epsilon) \geq 1 - \frac{1 - \mathrm{SC}(v)}{1 - \epsilon}, \tag{16}$$

which proves the desired lower bound. Hence, a higher $\mathrm{SC}(v)$ implies a larger lower bound on clustering clarity.

B Proof of theorem 2

NetLSD provides a compact and permutation-invariant representation of each cluster's spectral distribution, encoding cluster-level topology and enabling robust cluster comparison across views. In particular, the cluster topology is captured through the Laplacian spectrum and its multi-scale heat traces. We first introduce **Definition B.1** and **Lemma B.2**, which are the core of the theorem proof.

Definition B.1 (NetLSD Tsitsulin et al. [2018])**.** Let $\mathcal{H} = (V, E)$ be an undirected weighted graph with the symmetric normalized Laplacian L. Denote the eigenvalues of L by $\{\lambda_i\}_{i=1}^n$, where $n = |V|$. The heat trace at scale $\beta > 0$ is defined as

$$h_{\mathcal{H}}(\beta) = \operatorname{tr}(e^{-\beta L}) = \sum_{i=1}^{n} e^{-\beta\lambda_i}. \tag{17}$$

Given scales $\{\beta_1, \ldots, \beta_m\}$, the topology signature of $\mathcal{H}$ is defined by

$$\mathbf{s}(\mathcal{H}) = [\,\bar{h}(\beta_1), \ldots, \bar{h}(\beta_m)\,]^\top, \quad \bar{h}(\beta) := \frac{1}{n} h_{\mathcal{H}}(\beta). \tag{18}$$

Lemma B.2 (Kantorovich–Rubinstein duality Villani [2009])**.** For any probability measures μ and ν on $(\mathbb{R}_+, |\cdot|)$,

$$W_1(\mu, \nu) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \Big(\mathbb{E}_{x\sim\mu}[f(x)] - \mathbb{E}_{x\sim\nu}[f(x)] \Big). \tag{19}$$

where $W_1(\mu, \nu)$ is the first Wasserstein distance between the empirical spectral distributions μ and ν, which represents the minimum cost of transforming one spectral distribution into the other by shifting eigenvalues.

Proof: Let L be a symmetric Laplacian with eigenvalues $\{\lambda_i\}_{i=1}^n \subset [0, 2]$. Define the empirical spectral distribution as a probability measure on $\mathbb{R}_+$ by $\mu_L := \frac{1}{n}\sum_{i=1}^n \delta_{\lambda_i}$, where δ_x denotes the Dirac measure at x. For any measurable function $g : \mathbb{R}_+ \to \mathbb{R}$, the expectation under μ_L equals the discrete average

$$\mathbb{E}_{\lambda\sim\mu_L}[g(\lambda)] = \int g(\lambda)\, \mathrm{d}\mu_L(\lambda) = \frac{1}{n}\sum_{i=1}^{n} g(\lambda_i), \tag{20}$$

where λ is a random variable distributed according to μ_L. Taking $g(\lambda) = e^{-\beta\lambda}$ yields

$$\bar{h}(\beta) = \frac{1}{n}\sum_{i=1}^{n} e^{-\beta\lambda_i} = \mathbb{E}_{\lambda\sim\mu_L}\left[e^{-\beta\lambda}\right]. \tag{21}$$

Therefore, for each scale β_ℓ,

$$[\mathbf{s}(\mathcal{H})]_\ell = \bar{h}(\beta_\ell) = \mathbb{E}_{\lambda\sim\mu_L}\left[e^{-\beta_\ell\lambda}\right], \qquad \ell = 1, \ldots, m. \tag{22}$$

Thus, each entry $[\mathbf{s}(\mathcal{H})]_\ell$ depends only on μ_L, which is invariant under node relabeling. Moreover, $\mathbf{s}(\mathcal{H}) \in \mathbb{R}^m$ is an m-dimensional vector. Therefore, $\mathbf{s}(\mathcal{H})$ provides a compact and permutation-invariant summary of each cluster.

And then, consider two cluster-wise graphs $\mathcal{H}$ and $\mathcal{H}'$ with empirical spectral measures μ_L and $\mu_{L'}$. Fix a scale $\beta_\ell > 0$ and define the function $f_{\beta_\ell}(\lambda) = e^{-\beta_\ell\lambda}$. Since $|f'_{\beta_\ell}(\lambda)| = \beta_\ell e^{-\beta_\ell\lambda} \le \beta_\ell$ for all $\lambda \in \mathbb{R}_+$, the function f_{β_ℓ} is β_ℓ-Lipschitz. By Lemma B.2, for any K-Lipschitz function f on $\mathbb{R}_+$, setting $g = f/K$ gives $\|g\|_{\mathrm{Lip}} \le 1$ and thus

$$\Big|\mathbb{E}_{\lambda\sim\mu}[f(\lambda)] - \mathbb{E}_{\lambda\sim\nu}[f(\lambda)]\Big| \le K\, W_1(\mu, \nu). \tag{23}$$

Applying the above bound with $f = f_{\beta_\ell}$ yields

$$\big|[\mathbf{s}(\mathcal{H})]_\ell - [\mathbf{s}(\mathcal{H}')]_\ell\big| = \Big|\mathbb{E}_{\lambda\sim\mu_L}[f_{\beta_\ell}] - \mathbb{E}_{\lambda\sim\mu_{L'}}[f_{\beta_\ell}]\Big| \le \beta_\ell\, W_1(\mu_L, \mu_{L'}). \tag{24}$$

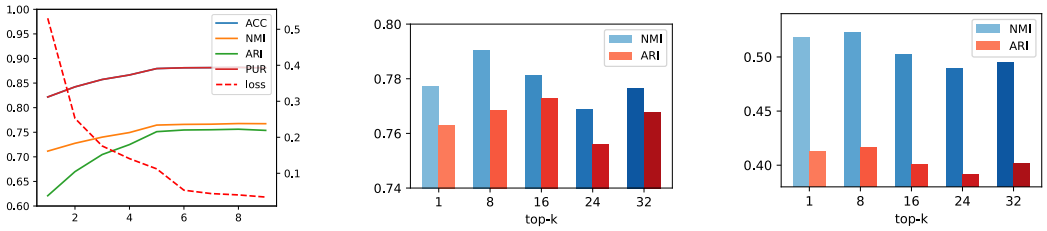
Stacking the bounds over $\ell = 1, \ldots, m$ gives

$$\|\mathbf{s}(\mathcal{H}) - \mathbf{s}(\mathcal{H}')\|_2 \le W_1(\mu_L, \mu_{L'}) \sqrt{\sum_{\ell=1}^{m} \beta_\ell^2}. \tag{25}$$

The above bound shows that the difference between two topology signatures is upper-bounded by a constant times $W_1(\mu_L, \mu_{L'})$. Thus, small changes in the spectral distribution only lead to small changes in the topology signature, making cross-view comparison based on $\mathbf{s}(\cdot)$ robust to minor perturbations.

C Additional Experimental Results

C.1 Convergence Analysis and Top-k Sensitivity



(a) Cifar10 convergence analysis (b) Caltech7-5 top-k sensitivity (c) Hdigit top-k sensitivity

Figure 4: Convergence analysis on Cifar10 and sensitivity analysis on Caltech7-5 and Hdigit.

Convergence analysis. Figure 4(a) shows the convergence curves on Cifar10. The loss drops quickly at the beginning and then gradually stabilizes. Meanwhile, the clustering metrics steadily improve and remain stable after convergence, which verifies the optimization stability of Fed-AGA.

Sensitivity analysis to the top-k parameter. Figure 4(b)(c) reports the sensitivity of the top-k attention mechanism on Caltech7-5 and Hdigit. Using $k = 1$ gives inferior performance, indicating that one-to-one selection is more vulnerable to mismatches. As k increases to a moderate range, the results improve significantly and achieve the best or near-best performance around $k = 8$. When k becomes too large, the performance drops due to the introduction of low-attention noisy correspondences. Overall, the method is not particularly sensitive to k within a reasonable range.

C.2 Reference-View Selection Analysis

Table 5: Reference-view analysis on Caltech7-5 and Hdigit.

Caltech7-5						Hdigit					
Reference view	**SC**	**ACC**	**NMI**	**ARI**	**PUR**	**Reference view**	**SC**	**ACC**	**NMI**	**ARI**	**PUR**
0	0.204	0.4743	0.3825	0.3087	0.5214	0	**0.169**	**0.5883**	0.5224	**0.4168**	**0.6303**
1	0.169	0.5557	0.3483	0.2683	0.5557	1	0.150	0.5608	**0.5324**	0.3971	0.5805
2	**0.288**	**0.8807**	**0.7902**	**0.7683**	**0.8807**	-	-	-	-	-	-
3	0.185	0.6600	0.5527	0.4638	0.6800	-	-	-	-	-	-
4	0.251	0.6621	0.5418	0.4574	0.6907	-	-	-	-	-	-

Table 5 reports the results obtained by treating each view as the reference in turn. On Caltech7-5, the view with the highest SC score also achieves the best final performance on all four metrics. On Hdigit, the highest-SC view gives the best or near-best overall performance. These results provide empirical support for selecting the highest-SC view as the reference.

C.3 Additional Ablation Studies

Table 6: Independent effects of the attention mechanism and the contrastive loss.

Attention	**Loss**	**Caltech7-5**				**Hdigit**			
		ACC	**NMI**	**ARI**	**PUR**	**ACC**	**NMI**	**ARI**	**PUR**
✓	Contrastive	**0.8807**	**0.7902**	**0.7683**	**0.8807**	**0.5883**	**0.5224**	**0.4168**	**0.6303**
✓	-	0.5221	0.4701	0.3551	0.5793	0.4261	0.3403	0.2181	0.4549
-	Contrastive	0.3529	0.1324	0.0939	0.3600	0.3009	0.1833	0.1112	0.3339
✓	MSE	0.5643	0.4142	0.3588	0.5929	0.3869	0.2539	0.1723	0.4090

Independent effects of attention and contrastive loss. Table 6 shows the independent effects of the attention mechanism and the contrastive loss. Removing either component causes a clear performance drop on both Caltech7-5 and Hdigit. Replacing the contrastive loss with MSE also

degrades the results. This confirms that both components are important for the fine alignment and fusion stage.

Table 7: Sensitivity analysis of the number of anchors M.

M	Caltech7-5				Hdigit			
	ACC	**NMI**	**ARI**	**PUR**	**ACC**	**NMI**	**ARI**	**PUR**
$1K$	0.8364	0.7163	0.6847	0.8364	0.4790	0.4379	0.3091	0.5226
$2K$	**0.8807**	**0.7902**	**0.7683**	**0.8807**	**0.5883**	**0.5224**	**0.4168**	**0.6303**
$4K$	0.8743	0.7802	0.7522	0.8743	0.5794	0.5188	0.4100	0.6219
$8K$	0.6843	0.5353	0.4404	0.6843	0.5625	0.5180	0.3937	0.6061

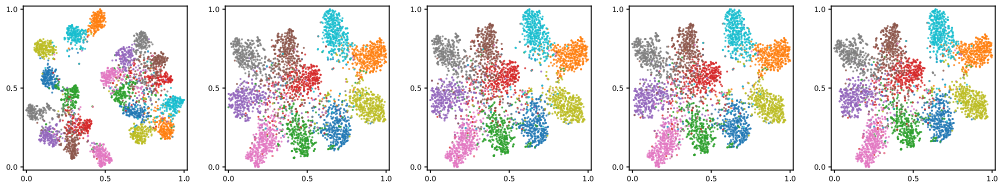
Sensitivity analysis to the anchor number M. Table 7 reports the sensitivity analysis of the anchor number M. Among the tested choices, $M = 2K$ achieves the best overall performance on both datasets, where K is the number of clusters. Therefore, we set $M = 2K$ by default.

Table 8: Ablation study of key components on Cifar10.

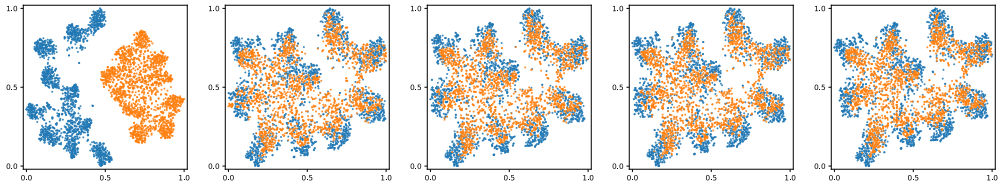
	$\mathcal{L}_{sp}$	$\mathbf{R}$	$\mathcal{L}_{pull}$	$\mathcal{FA}$	ACC	NMI	ARI	PUR
(A)	✓				0.3513	0.1760	0.1354	0.3581
(B)	✓	✓	✓		0.5250	0.2530	0.2269	0.5250
(C)	✓			✓	0.8197	0.6975	0.6235	0.8197
(D)	✓		✓	✓	0.8242	0.6967	0.6432	0.8242
(E)	✓	✓		✓	0.8337	0.7245	0.6332	0.8337
(F)	✓	✓	✓	✓	**0.8809**	**0.7674**	**0.7538**	**0.8809**

Additional component ablation on Cifar10. Table 8 reports additional component ablations on Cifar10. Introducing $\mathcal{FA}$ brings clear improvements over the settings without it, showing the importance of the fine alignment and fusion stage. Moreover, removing either $\mathcal{L}_{\text{pull}}$ or $\mathbf{R}$ degrades the performance, indicating that both are important for coarse alignment. Overall, the coarse alignment components and the fine alignment and fusion stage jointly improve clustering quality.

C.4 Additional Visualization of Coarse Alignment



(a) Colored by cluster, rounds 0, 2, 4, 6, 8



(b) Colored by view, rounds 0, 2, 4, 6, 8

Figure 5: Evolution of coarsely aligned anchor graphs $\mathbf{G}_t$ on Cifar10 at communication rounds 0, 2, 4, 6, and 8. (a) Points are colored by cluster labels; closer same-cluster samples across views indicate clearer cross-view cluster correspondence. (b) Points are colored by view labels; closer and more thoroughly mixed representations from different views indicate stronger cross-view consistency.

Figure 5 visualizes the evolution of coarse alignment on Cifar10. As the communication rounds proceed, same-category samples from different views become progressively closer, and the two views become more mixed. This shows that the proposed coarse alignment stage can gradually recover reliable cross-view correspondences.

C.5 Additional Computation Cost Analysis

Table 9: Host memory consumption comparison of traditional clustering-based methods on seven datasets, where Host Mem. denotes peak host memory usage.

Dataset	Metric	Methods				
		MVC-UM	OpVuC	TUMCR	UPMGC	Ours
Scene	Host Mem. (GB)	1.37	**0.16**	0.46	0.63	1.51
Caltech7-5	Host Mem. (GB)	0.45	**0.07**	0.42	0.19	1.57
Caltech102-5	Host Mem. (GB)	16.17	**0.87**	3.67	4.13	1.69
Hdigit	Host Mem. (GB)	3.11	**0.80**	1.67	2.37	1.73
Aloi	Host Mem. (GB)	13.95	**1.18**	4.22	4.39	1.59
NUSWIDEOBJ	Host Mem. (GB)	> 128.00	7.11	> 128.00	> 128.00	**1.86**
Cifar10	Host Mem. (GB)	> 128.00	20.28	> 128.00	> 128.00	**2.88**

Traditional clustering-based methods. Table 9 reports the peak host memory usage of traditional baselines across multiple datasets. Overall, our method achieves lower host memory consumption than most traditional approaches under the same experimental settings. Notably, on larger datasets such as NUSWIDEOBJ and Cifar10, several baselines exhibit sharply increased memory usage or even fail due to excessive memory demands, whereas our method remains stable with moderate memory usage, highlighting its advantage in scaling to large datasets.

Table 10: Computation cost comparison of deep federated clustering methods on seven datasets, where Time denotes wall-clock training time (s), and GPU Mem. and Host Mem. denote peak GPU memory and peak host memory usage, respectively.

Dataset	Metric	Methods		
		MGCD	MGCCFF	Ours
Scene	Time (s)	350.53	**102.11**	122.74
	GPU Mem. (GB)	0.74	2.58	**0.27**
	Host Mem. (GB)	2.71	**1.35**	1.51
Caltech7-5	Time (s)	102.92	110.51	**41.59**
	GPU Mem. (GB)	0.76	0.50	**0.40**
	Host Mem. (GB)	1.60	**1.42**	1.57
Caltech102-5	Time (s)	1062.50	687.06	**333.57**
	GPU Mem. (GB)	3.41	15.61	**0.90**
	Host Mem. (GB)	5.01	**1.50**	1.69
Hdigit	Time (s)	149.04	280.16	**75.74**
	GPU Mem. (GB)	1.43	9.50	**0.37**
	Host Mem. (GB)	2.87	**1.56**	1.73
Aloi	Time (s)	265.04	743.56	**242.42**
	GPU Mem. (GB)	1.43	18.07	**0.70**
	Host Mem. (GB)	2.89	**1.45**	1.59
NUSWIDEOBJ	Time (s)	4680.52	–	**1788.27**
	GPU Mem. (GB)	4.72	OOM	**0.99**
	Host Mem. (GB)	97.93	–	**1.86**
Cifar10	Time (s)	3371.08	–	**2376.29**
	GPU Mem. (GB)	4.84	OOM	**1.72**
	Host Mem. (GB)	59.05	–	**2.88**

Deep federated clustering methods. Table 10 compares deep federated baselines in terms of training time and memory usage. Our method achieves competitive training efficiency with notably lower GPU memory consumption, and remains feasible on the largest datasets where some baselines encounter out-of-memory issues. These results indicate that Fed-AGA is efficient and scalable under federated constraints, with lower computation cost.

C.6 Additional Privacy Leakage Analysis

Table 11: Additional privacy leakage results on the remaining datasets.

Dataset	Reconstruction MSE $\times 10^3 \uparrow$	
	FedDMVC	**Ours**
Scene	0.945	**2.233**
Caltech7-5	13.058	**15.670**
Caltech102-5	3.577	**4.429**
Hdigit	28.962	**44.653**
NUSWIDEOBJ	8.532	**11.245**

Table 11 extends the decoder-based reconstruction evaluation to the remaining datasets. Fed-AGA consistently yields higher reconstruction MSE than FedDMVC, suggesting that anchor graphs are harder to reconstruct into original features and indicating lower empirical privacy leakage risk. This privacy analysis focuses on a representative reconstruction-based leakage evaluation to assess empirical privacy leakage risk.

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: The main claims are stated in the abstract and introduction, and are supported by the proposed method, theoretical analysis, and experimental results.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We discuss the scope of the empirical privacy leakage analysis in Appendix C.6.

   Guidelines:

   - The answer [N/A] 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.

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: The paper provides formal theoretical statements and the corresponding proofs in Appendices A and B.

Guidelines:

- The answer [N/A] 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: The paper provides the necessary experimental settings, implementation details, datasets, metrics, and hardware information for reproducing the main results.

   Guidelines:

   - The answer [N/A] 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: The code of the proposed method is provided in the supplementary material, and the used datasets are public benchmark datasets.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the training details, model architecture, optimizer, hyperparameters, and evaluation protocol.

   Guidelines:

   - The answer [N/A] 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: [No]

   Justification: Almost all compared baselines do not include the statistical significance in experiments thus we do not report it.

   Guidelines:

   - The answer [N/A] 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.
   - 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: The paper reports the hardware configuration and provides runtime and memory comparisons.

   Guidelines:

   - The answer [N/A] 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: The research conducted in the paper conforms to the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: The paper does not involve applications with direct societal implications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not involve the description of safeguards for responsible release of data or models with a high risk for misuse.

    Guidelines:

    - The answer [N/A] 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: The code for the comparison methods in the experimental section all includes proper citations.

    Guidelines:

    - The answer [N/A] 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.
- 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: [Yes]

    Justification: We have provided the source code of our algorithm, which is included in the supplementary materials.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: The paper focuses on machine learning algorithm research and does not involve crowdsourcing or research with human subjects at all.

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: Our manuscript focuses on algorithmic research, and it does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] 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.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The method in this paper does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.