BEYOND VIEW HETEROGENEITY: FEDERATED INCOMPLETE MULTI-VIEW CLUSTERING WITH ARBITRARY VIEW SUBSETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated multi-view clustering (FedMVC) learns cluster structures from multi-view data distributed across clients without centralizing raw data. Beyond differences in view-specific feature spaces and client view subsets, individual samples may lack views available at their client. Client-level view heterogeneity and sample-level view missingness jointly determine the cross-view supervision and relational evidence available for learning common representations. In this paper, we propose Federated Incomplete Multi-view Clustering via Prediction and Alignment (FIMCPA) for this more general setting. Specifically, masked cross-view predictive pretraining exploits partial joint observations to capture shared information for common representation initialization. Relation-verified contrastive alignment balances sample contributions and uses shared-view agreement to mitigate potential false-negative repulsion, promoting consistency among common and observed-view representations. Observation-weighted aggregation further accounts for differences in available data during model aggregation. These designs allow clients with different observation patterns to contribute to common representation learning without imputing genuinely missing views. Theoretical analysis provides insight into relation verification and its effects on the contrastive objective. Extensive experiments demonstrate superior clustering performance, with further analyses supporting the effectiveness of FIMCPA under varying observation conditions and client compositions. Code is available at `https://anonymous.4open.science/r/FIMCPA-9FB9/`.

1 INTRODUCTION

Federated multi-view clustering (FedMVC) learns cluster structures from multi-view data distributed across clients while keeping raw data local (Huang et al., 2022; Chen et al., 2023). In practice, view heterogeneity arises from differences in both view-specific feature spaces and the view subsets available across clients. For example, CT images, laboratory measurements, and clinical text provide different feature representations, while hospitals may offer different sets of examinations, resulting in *client-level view heterogeneity*. Within each hospital, individual patients may undergo only some of the available examinations, leading to *sample-level view missingness*. Together, client view availability and sample-level view missingness determine the actual observations available for collaborative clustering.

Recent FedMVC methods accommodate increasingly diverse client view configurations. FMCSC supports single- and full-view clients (Chen et al., 2024), while EFDMVC (Wei et al., 2026) and AFMCC (Xu et al., 2026) accommodate more general client view subsets. Client view configurations, however, do not specify which views are jointly observed for each sample, and hence which cross-view information is available for learning. We therefore go beyond client-level view heterogeneity by jointly modeling client view availability and sample-level view missingness. This formulation includes complete observations and sample-level view missingness under uniform or heterogeneous client view configurations.

Client-level view heterogeneity and sample-level view missingness jointly limit the cross-view supervision available for learning common representations. Client view subsets restrict local view

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

pairings, while sample-level view missingness determines which samples retain them. Even without the full view set, co-observed views provide predictive supervision: one observed view can serve as a target for the others. Within-view reconstruction alone does not explicitly require such cross-view prediction. The challenge is to exploit this supervision to initialize common representations using only observed views as targets.

The same observation differences also affect contrastive learning and model aggregation. Samples with more observed views supply more representation pairs and can dominate an equally weighted pairwise objective. Meanwhile, instance-level contrastive learning can repel samples from the same unknown cluster (Chuang et al., 2020; Huynh et al., 2022). Cross-view evidence for screening potential false negatives depends on the views jointly observed for each sample pair. For example, samples with view subsets $\{1, 2\}$ and $\{2, 3\}$ share only one view despite each having two. Likewise, acquiring a view does not specify how many local samples support its model update. Contrastive learning and model aggregation must therefore account for actual observations, rather than client view configurations alone.

To address these challenges, we propose Federated Incomplete Multi-view Clustering via Prediction and Alignment (FIMCPA), which learns common representations by using actual observations at the sample, sample-pair, and model-update levels. Figure 1 illustrates the framework. Specifically, Masked Cross-view Predictive Representation (MCPR) pretraining predicts held-out observed views from the remaining observations, encouraging cross-view information sharing beyond within-view reconstruction. Subsequent contrastive learning aligns common and observed-view representations while balancing each sample's contribution. To screen potential false negatives, relation verification requires mutual neighborhood agreement in every shared observed view, with at least two shared views. Verified pairs are excluded from negative comparisons without introducing additional positives. Throughout training, observation-weighted aggregation combines view-specific model blocks according to their observed support. Together, these designs exploit available cross-view information while accounting for the different observations supporting local learning and federated collaboration.

Our main contributions are as follows:

1. We propose FIMCPA, a federated multi-view clustering method that jointly handles client-level view heterogeneity and sample-level view missingness, extending collaborative clustering to a more general incomplete multi-view setting.
2. We design masked cross-view predictive pretraining and relation-verified contrastive alignment, together with observation-weighted aggregation, to learn common representations from available observations and explore cluster structures in incomplete multi-view data distributed across heterogeneous clients.
3. Theoretical analysis and extensive experiments support the effectiveness of FIMCPA, demonstrating its superior clustering performance and adaptability to varying observation conditions and client compositions.

2 Related Work

2.1 Federated Incomplete Multi-View Clustering

Federated incomplete multi-view clustering explores cluster structures in distributed multi-view data without centralizing raw data. Existing studies address incomplete observations under different collaboration settings. FedDMVC investigates feature heterogeneity and incompleteness (Chen et al., 2023), while FCUIF addresses unaligned and incomplete data fusion (Ren et al., 2024). FIMCFG and EFIMVC further exploit global information to improve clustering with missing views (Chao et al., 2025; Liu et al., 2025). ESFMC studies the balance between information sharing and privacy protection and extends its framework to incomplete scenarios (Chen et al., 2025). HFMVC addresses heterogeneous representations through cross-client alignment of corresponding samples and selective aggregation among clients holding the same view (Jiang et al., 2024).

Client-level view heterogeneity has also motivated methods that accommodate different view subsets across clients. FMCSC enables collaboration between single-view and full-view clients (Chen

et al., 2024), while H2-FedMVC further investigates heterogeneous clients under category and quantity skew (Wang et al., 2026). EFDMVC and AFMCC accommodate more general client view configurations (Wei et al., 2026; Xu et al., 2026). Although these methods broaden the applicability of FedMVC, client view configurations alone do not determine which views are observed for each sample. Sample-level view missingness further changes the cross-view information supporting representation learning and federated collaboration. FIMCPA addresses this setting by learning common representations from actual observations across heterogeneous clients.

2.2 Self-Supervised Multi-View Representation Learning

Self-supervised multi-view learning uses cross-view prediction and alignment to learn shared representations. MFLVC separates reconstruction and cross-view contrastive objectives across feature levels to reconcile view-specific information with common semantics (Xu et al., 2022). Masked multimodal autoencoding learns representations through prediction across masked patches and modalities (Bachmann et al., 2022). For incomplete multi-view clustering, COMPLETER combines contrastive learning with cross-view prediction to recover missing views (Lin et al., 2021). HSACC jointly explores semantic alignment and cooperative completion, whereas FreeCSL learns consensus semantics without imputation or explicit cross-view alignment (Ding et al., 2025; Dai et al., 2025). In FIMCPA, prediction initializes common representations: each target is an observed view withheld from the input, and subsequent contrastive training uses only actual observations.

Contrastive learning depends on reliability at both the view-pair and sample-pair levels. SEM downweights unreliable view pairs in its contrastive loss and regularizes representations through reconstruction (Xu et al., 2023). At the sample-pair level, debiased contrastive learning and false negative cancellation address the adverse effects of treating semantically related samples as negatives (Chuang et al., 2020; Huynh et al., 2022). In multi-view clustering, SURE studies robust learning under incomplete information (Yang et al., 2023), while DIVIDE and CANDY refine sample relations to improve contrastive learning (Lu et al., 2024; Guo et al., 2024). Under sample-level view missingness, the shared observations available for assessing relations vary across sample pairs. FIMCPA uses agreement across these observations to screen potential false negatives and learn common representations for clustering.

3 Problem Setting

Let C clients hold non-overlapping sample sets $\mathcal{I}_c$, with $N_c = |\mathcal{I}_c|$, and refer to a common registry of V views. Binary acquisition masks A_{cv} encode client view availability; their variation across clients represents client-level view heterogeneity. Binary conditional observation masks M_{civ} encode sample-level view missingness. The realized support is

$$R_{civ} = A_{cv} M_{civ}. \tag{1}$$

Here $A_{cv} = 0$ means that view v lies outside client c's acquisition support; conditional on $A_{cv} = 1$, $M_{civ} = 0$ means that this view is missing for sample i. The value of M_{civ} is immaterial when $A_{cv} = 0$; we define sample-level missingness rates only over entries with $A_{cv} = 1$. Client c may use $\mathbf{x}_i^v$ only when $R_{civ} = 1$, and every sample retains at least one observed view, $\sum_v R_{civ} \geq 1$.

Clients are called full-, partial-, or single-view clients if they acquire all V views, more than one but fewer than V views, or exactly one view, respectively. These roles do not determine the observations of each sample. Learning and aggregation instead use the observed count $n_{cv} = \sum_{i \in \mathcal{I}_c} R_{civ}$ and the shared-view set $\mathcal{S}_{cij} = \{v : R_{civ} R_{cjv} = 1\}$ for each sample pair. Neither is recoverable in general from $(A_{c:}, N_c)$ alone (Appendix A.1).

The goal is to learn a shared model producing comparable representations $\mathbf{h}_i$ from realized observations and obtain K clusters without transmitting raw samples or per-sample embeddings. We assume that the number of clusters, K, is known. Here, "arbitrary view subsets" refers to the admissible acquisition and observation masks under the nonempty-sample constraint. Recovering common cluster semantics also requires informative cross-view associations; the formulation does not guarantee identifiability for disconnected view groups without such information. Both training stages use the same support constraints; MCPR holds out observed targets, but neither stage reconstructs genuinely missing views. Appendices D and C give the complete notation and benchmark construction.

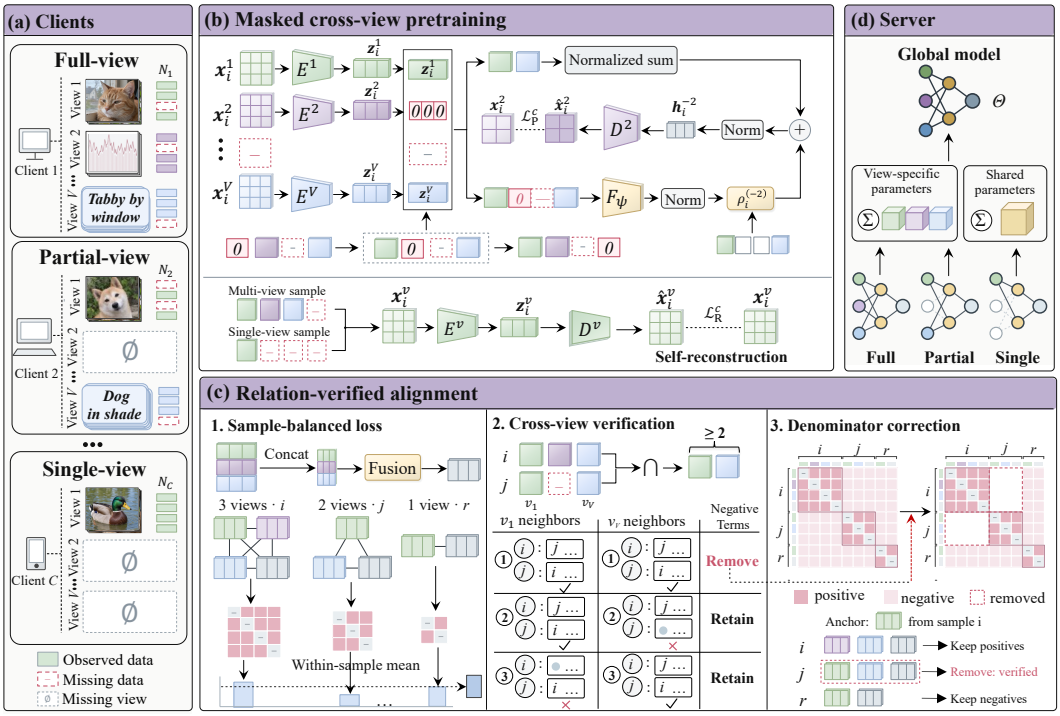


Figure 1: The framework of FIMCPA. Clients have different available view subsets, and individual samples may lack some locally available views. Masked cross-view predictive pretraining initializes common representations from available observations, followed by relation-verified contrastive alignment to explore cluster structures. The server performs observation-weighted aggregation and distributes the updated models to clients, enabling collaborative clustering under heterogeneous and incomplete observations.

4 Methodology

4.1 Overview

FIMCPA addresses client-level view heterogeneity and sample-level view missingness through prediction and alignment. Predictive pretraining supplies cross-view supervision beyond reconstruction. Contrastive alignment promotes cross-view consistency while addressing unequal sample contributions and potential false negatives. Both stages share an autoencoder-based model (Chen et al., 2024; Wei et al., 2026) and use observation-weighted aggregation before federated clustering (Figure 1). Because clients hold non-overlapping sample sets, sample-wise variables omit the redundant client subscript; support, losses, and model updates retain c.

4.2 Masked cross-view predictive pretraining

Common representation from observed views. Both learning stages use a view-specific encoder–decoder pair $(E^v_{\phi_v}, D^v_{\theta_v})$ for each registered view and a fusion network $F_\psi : \mathbb{R}^{Vd} \to \mathbb{R}^d$. An observed input $\mathbf{x}_i^v \in \mathbb{R}^{D_v}$ is encoded as $\mathbf{z}_i^v = E^v_{\phi_v}(\mathbf{x}_i^v) \in \mathbb{R}^d$. Local objectives omit the client subscript on model parameters for brevity. For minibatch $\mathcal{B}_c$, write $\mathcal{B}_{cv} = \{i \in \mathcal{B}_c : R_{civ} = 1\}$. Both stages use the observed-view reconstruction loss

$$\mathcal{L}_{\mathrm{R}}^c = \sum_{v:|\mathcal{B}_{cv}|>0} \sum_{i\in\mathcal{B}_{cv}} \frac{1}{|\mathcal{B}_{cv}|} \left\| \mathbf{x}_i^v - D^v_{\theta_v}\left(E^v_{\phi_v}(\mathbf{x}_i^v)\right) \right\|_2^2 . \tag{2}$$

Only observed sample–view pairs contribute; implementation normalization is specified in Appendix B.

Write $\mathrm{norm}(\mathbf{a})$ for ℓ_2 normalization with an ϵ clamp (Appendix B). The fusion branch concatenates view representations, with $\mathbf{z}_i^v = \mathbf{0}$ for unobserved views: $\widehat{\mathbf{u}}_i = \mathrm{norm}\big(F_\psi([R_{ci1}\mathbf{z}_i^1; \ldots; R_{ciV}\mathbf{z}_i^V])\big)$. To retain a direct contribution from observed evidence, we combine the normalized fusion output with the normalized sum $\mathbf{b}_i = \mathrm{norm}(\sum_v R_{civ}\,\mathrm{norm}(\mathbf{z}_i^v))$. With realized support fraction $\rho_i = V^{-1}\sum_v R_{civ}$, the common representation is

$$\mathbf{h}_i = \mathrm{norm}(\mathbf{b}_i + \rho_i \widehat{\mathbf{u}}_i)\,. \tag{3}$$

The observed-view sum provides a direct path from the available latent representations; ρ_i scales the nonlinear branch by the fraction of registered views observed for that sample. This coefficient depends on support size and does not estimate view quality. Appendix A.2 characterizes the support-only weighting rule and its spherical interpretation.

Cross-view predictive supervision. Within-view reconstruction does not directly constrain the common representation through cross-view prediction. MCPR introduces this constraint by predicting an observed target from the remaining observations, encouraging the common representation to capture shared information before contrastive alignment. Eligible samples are $\mathcal{E}_{cv} = \{i \in \mathcal{B}_c : R_{civ} = 1,\ \sum_u R_{ciu} \geq 2\}$. For each target v, we zero its slot for these samples and recompute the fusion branch, observed-view sum, and support fraction from the remaining views, giving $\mathbf{h}_i^{(-v)}$ with $\rho_i^{(-v)} = V^{-1}\sum_{u\neq v} R_{ciu}$. The existing target-view decoder yields the prediction loss

$$\mathcal{L}_{\mathrm{P}}^c = \sum_{v:|\mathcal{E}_{cv}|>0} \sum_{i\in\mathcal{E}_{cv}} \frac{1}{|\mathcal{E}_{cv}|} \left\| \mathbf{x}_i^v - D_{\theta_v}^v(\mathbf{h}_i^{(-v)}) \right\|_2^2 . \tag{4}$$

Every eligible target participates, but no target enters its own input representation and no genuinely missing view supplies supervision. Pretraining minimizes $\mathcal{L}_{\mathrm{pre}}^c = \frac{1}{2}(\mathcal{L}_{\mathrm{R}}^c + \mathcal{L}_{\mathrm{P}}^c)$ when the minibatch contains eligible prediction targets, and uses $\mathcal{L}_{\mathrm{R}}^c$ otherwise. Single-view samples contribute only to reconstruction. The aggregated pretrained model initializes contrastive training. MCPR adds no trainable parameters, but requires extra local computation (Appendix B.4).

The target is removed from both the observed-view sum and the fusion input to prevent its information from reaching the predictor through the residual path. Contrastive training restores the original observation masks and uses no predicted views.

4.3 Relation-verified contrastive alignment

Contrastive alignment must address unequal sample contributions within the objective and potentially false-negative relations between samples. We therefore combine sample-balanced alignment with cross-view relation verification.

Sample-balanced alignment. Each sample's common and observed-view representations form a fully connected positive set, called an observed semantic clique, containing its common node $\mathbf{h}_i^0 = \mathbf{h}_i$ and observed-view nodes $\mathbf{h}_i^v = \mathbf{z}_i^v$, normalized when computing similarities. Let $\mathcal{U}_c$ collect these minibatch nodes, indexed by (i, a), and let $\mathcal{P}_{ci}^a$ contain all other nodes of sample i. Every anchor has a positive because every sample retains an observed view. A sample with s observed views supplies $s + 1$ nodes and $s(s + 1)$ ordered positive pairs. Averaging all such pairs equally would therefore give greater objective weight to samples with more observed views.

Cross-view relation verification. Instance-level contrastive learning treats different samples as negatives, including pairs from the same unknown cluster. We screen pairs for consistent neighborhood evidence in their shared observations before deciding which negative terms to retain. The test controls denominator membership; it does not assign cluster labels to the pair. For a view with $\nu_{cv} = |\mathcal{B}_{cv}| \geq 2$, set

$$q_{cv} = \max\Big(1, \min\Big(\nu_{cv} - 1, \Big\lfloor \frac{\nu_{cv}}{K} \Big\rfloor\Big)\Big)\,. \tag{5}$$

For $i \notin \mathcal{B}_{cv}$ or $\nu_{cv} < 2$, set $\mathcal{N}_{cv}(i) = \varnothing$; otherwise, let $\mathcal{N}_{cv}(i)$ contain the q_{cv} observed samples with the highest cosine similarity to sample i, excluding i itself. For $i \neq j$, the symmetric verifier is

$$G_{cij} = \mathbf{1}[|\mathcal{S}_{cij}| \geq 2] \prod_{v\in\mathcal{S}_{cij}} \mathbf{1}[j \in \mathcal{N}_{cv}(i)]\mathbf{1}[i \in \mathcal{N}_{cv}(j)]. \tag{6}$$

Set $G_{cii} = 0$; $\mathbf{1}[\cdot]$ denotes the indicator function. Neighborhoods are computed without propagating gradients through neighbor selection; views with fewer than two observed samples cannot verify a pair. The neighborhood-size rule assumes approximately balanced clusters and does not use a similarity threshold. It can under-cover majority clusters and over-cover minority clusters. Mutual agreement is required in every shared view, and the two-view minimum excludes verification on single-view clients. Verification adds no trainable parameters or extra communication.

The test uses each pair's shared-view set. The pair with observed-view subsets $\{1, 2\}$ and $\{2, 3\}$ in Section 1 therefore retains its negative terms because only one view is shared. For an eligible pair, disagreement in any shared view also retains repulsion. Passing this unsupervised test is evidence used by the algorithm, not a guarantee of shared cluster membership. Appendix A.6 gives conditions under which shared-view agreement increases relation precision.

One-sided contrastive correction. For anchor (i, a), we remove all nodes of other samples verified as related to sample i from the denominator, while preserving $\mathcal{P}^a_{ci}$. Excluding the anchor itself gives the candidate set

$$\mathcal{A}^a_{ci} = \mathcal{U}_c \setminus \left(\{(i, a)\} \cup \{(j, r) \in \mathcal{U}_c : j \neq i,\ G_{cij} = 1\}\right). \tag{7}$$

All within-sample positives remain in this set, so $\mathcal{P}^a_{ci} \subseteq \mathcal{A}^a_{ci}$ (Proposition 3). The anchor loss uses the multi-positive form studied in supervised contrastive learning (Khosla et al., 2020); here, positives are representations of the same sample and require no class labels. With cosine similarity sim and temperature $\tau > 0$, the anchor loss averages over these positives:

$$\ell^a_{ci} = -\frac{1}{|\mathcal{P}^a_{ci}|} \sum_{(i,b) \in \mathcal{P}^a_{ci}} \log \frac{\exp\big(\mathrm{sim}(\mathbf{h}^a_i, \mathbf{h}^b_i)/\tau\big)}{\sum_{(j,r) \in \mathcal{A}^a_{ci}} \exp\big(\mathrm{sim}(\mathbf{h}^a_i, \mathbf{h}^r_j)/\tau\big)}. \tag{8}$$

Incorrect removal can suppress useful repulsion and alter retained gradients through denominator renormalization. Appendices A.7 and A.8 analyze the resulting loss and single-anchor update perturbations, with fixed candidates and centers for the latter. The anchor loss averages positives, but samples still supply different numbers of anchors. We therefore average over the $1 + \sum_v R_{civ}$ anchors of each sample before averaging across the minibatch, giving each sample equal outer weight:

$$\mathcal{L}^c_{\mathrm{C}} = \frac{1}{|\mathcal{B}_c|} \sum_{i \in \mathcal{B}_c} \frac{1}{1 + \sum_v R_{civ}} \sum_{a:\,(i,a) \in \mathcal{U}_c} \ell^a_{ci}. \tag{9}$$

We combine this sample-balanced alignment objective with reconstruction to preserve view-specific information during contrastive training:

$$\mathcal{L}^c = \mathcal{L}^c_{\mathrm{R}} + \alpha \mathcal{L}^c_{\mathrm{C}}, \tag{10}$$

where α balances contrastive alignment against reconstruction. This stage uses the original realized masks and no predictive term. Single-view clients use the same objective without cross-sample verification.

4.4 Federated optimization and clustering

Observation-weighted aggregation. We follow the iterative model-averaging framework of FedAvg (McMahan et al., 2017), with block-specific weights based on observed support. Local sample count does not indicate how many observations contribute to each view-specific model. The encoder–decoder for view v and the corresponding columns of the fusion input matrix are therefore aggregated with weight n_{cv}, while the shared fusion parameters are aggregated with weight $m_c = \sum_v n_{cv}$. For a local parameter block $\boldsymbol{\vartheta}_c$ with weight q_c, the server applies the following update when $\sum_c q_c > 0$:

$$\overline{\boldsymbol{\vartheta}} = \frac{\sum_c q_c \boldsymbol{\vartheta}_c}{\sum_c q_c}. \tag{11}$$

This rule excludes clients with no observations of a view from that view-specific average. The weights measure observed support rather than nonzero gradients: single-view clients still contribute to shared-block averaging during MCPR, although they supply no predictive gradient to the fusion network. Appendix A.9 derives the update as a weighted parameter barycenter, not an optimum of the training risk.

Federated clustering. Clients encode their samples with the final global model and jointly fit K global cluster centers without uploading embeddings. Each client runs local K_c-means with $K_c = \min(K, N_c)$ and sends centers with membership counts; weighted server-side K-means supplies initial global cluster centers. In each federated Lloyd iteration, clients assign samples to the broadcast centers and return cluster sums $\mathbf{S}_{ck}$ and counts n_{ck}. The server updates each nonempty cluster by

$$\boldsymbol{\mu}_k \leftarrow \frac{\sum_c \mathbf{S}_{ck}}{\sum_c n_{ck}}. \tag{12}$$

Repeating broadcasts and local assignments implements centralized Lloyd refinement in exact arithmetic under the matching conditions in Appendix B.1. After the final center broadcast, each client obtains its cluster labels as $\widehat{k}_{ci} = \arg\min_{k\in[K]} \|\mathbf{h}_i - \boldsymbol{\mu}_k\|_2^2$. Algorithm 1 gives the full execution flow.

5 Experiments

5.1 Experimental Settings

Datasets and observation protocol. We evaluate eight benchmark configurations: Caltech-5V, UCI, HW, OutdoorScene, CCV, Fashion, CIFAR10-MV, and MNIST-3V, covering 1.4K–60K samples. These include configurations reported in FMCSC, ESFMC, and EFIMVC (Chen et al., 2024; 2025; Liu et al., 2025). UCI and HW use different view subsets of the same source collection; MNIST-3V is the 60,000-sample, three-view feature version of MNIST. Appendix C.1 gives feature dimensions and source records. Unless otherwise stated, we use class-stratified client partitions with a 2:1:1 ratio of full-, partial-, and single-view clients. Half of the samples on clients with at least two acquired views are made incomplete, with at least one view retained. The rate of incomplete samples differs from the fraction of missing view entries. All methods use the same partitions, observation masks, and five paired seeds. We evaluate clustering accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (ARI).

Comparison methods and training. We compare with three centralized incomplete multi-view clustering (IMVC) baselines: DSIMVC (Tang & Liu, 2022), FreeCSL (Dai et al., 2025), and HSACC (Ding et al., 2025). For centralized methods, we concatenate samples across clients while retaining their observation masks. We also compare with four federated baselines: FMCSC (Chen et al., 2024), AFMCC (Xu et al., 2026), EFDMVC (Wei et al., 2026), and H2-FedMVC (Wang et al., 2026). For FIMCPA, we use a batch size of 256 and set $\alpha = \tau = 0.5$ across all eight benchmark configurations.

5.2 Results and Analysis

Clustering Results. Table 1 presents a quantitative comparison under joint client-level view heterogeneity and sample-level view missingness. Each experiment is independently conducted five times, and the average results are reported. Remarkably, FIMCPA achieves exceptional performance across all eight datasets, surpassing the compared centralized and federated methods. These results further support the overall design of FIMCPA, which learns common representations from the available observations across clients to effectively explore cluster structures in distributed incomplete multi-view data.

Ablation Studies. Table 2 compares four variants without prediction, target masking, alignment, or verification. Reconstruction pretraining retains the same schedule, and removing verification preserves positives and their weights.

Full achieves the best mean performance across all four datasets. Removing prediction reduces performance despite retaining reconstruction pretraining, supporting cross-view predictive supervision for common representation initialization. Removing target masking also degrades performance while retaining the prediction objective and exposing its target as input. Together, these observations support learning shared information by predicting an observed view from the remaining observations, rather than merely adding a prediction objective.

Table 1: Clustering performance (%) on eight datasets, averaged over five runs. Best and second-best means are bold and underlined.

Method	Caltech-5V			UCI			HW			OutdoorScene		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
DSIMVC	45.76	36.40	26.10	62.82	61.18	46.87	57.62	54.19	40.07	49.21	39.92	27.94
FreeCSL	79.36	66.61	62.40	74.19	71.59	62.25	82.57	77.70	71.98	53.45	46.85	33.54
HSACC	40.96	41.64	19.34	66.89	70.53	51.41	65.81	69.32	47.33	46.97	45.24	28.15
FMCSC	71.26	53.15	49.47	75.56	66.88	60.44	77.29	68.82	63.38	60.36	43.94	36.49
AFMCC	71.49	55.92	51.45	77.42	70.75	63.52	76.83	68.30	61.41	68.39	50.27	43.43
EFDMVC	65.67	52.04	44.77	76.04	71.24	63.12	80.53	73.86	68.90	61.38	48.20	37.88
H2-FedMVC	78.66	60.39	59.28	83.75	73.00	68.97	84.71	71.86	69.57	70.98	53.54	47.01
FIMCPA	**79.51**	**66.94**	**63.65**	**85.92**	**76.58**	**72.96**	**90.20**	**81.72**	**80.19**	**73.90**	**57.42**	**50.85**

Method	CCV			Fashion			CIFAR10-MV			MNIST-3V		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
DSIMVC	17.50	14.76	5.55	74.08	72.18	62.78	50.43	43.89	29.51	96.43	93.47	93.63
FreeCSL	20.74	17.01	7.61	87.03	81.27	77.53	61.56	53.93	42.55	98.66	96.07	97.03
HSACC	14.91	11.00	3.11	81.07	78.10	70.70	37.09	34.54	9.01	61.99	70.68	37.29
FMCSC	18.10	13.49	5.38	83.44	73.00	69.62	49.20	37.85	27.95	93.18	85.65	85.77
AFMCC	19.70	15.95	6.56	88.75	80.80	78.23	58.05	46.61	37.84	98.49	95.65	96.68
EFDMVC	20.56	16.95	7.25	84.22	76.75	71.38	51.60	42.86	31.09	98.27	95.12	96.22
H2-FedMVC	16.62	12.67	4.80	89.56	81.91	79.72	61.58	46.10	38.09	98.39	95.33	96.48
FIMCPA	**24.45**	**18.78**	**8.77**	**89.81**	**82.33**	**79.78**	**66.70**	**54.61**	**45.01**	**99.00**	**96.93**	**97.78**

Table 2: Ablation studies of FIMCPA on four datasets with $r = 0.5$.

Variant	UCI			HW			OutdoorScene			CIFAR10-MV		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
w/o prediction	77.25	67.72	61.04	89.32	80.36	78.47	72.78	56.23	49.65	65.76	54.30	43.14
w/o target masking	81.86	74.55	69.06	89.05	81.20	79.30	72.65	56.12	49.21	66.30	54.33	44.22
w/o alignment	84.19	74.55	70.09	67.28	59.37	50.80	63.26	46.41	38.66	56.16	46.76	34.61
w/o verification	85.61	76.10	72.46	88.34	80.02	77.84	73.10	56.29	49.58	60.70	49.25	38.26
FIMCPA (Full)	**85.92**	**76.58**	**72.96**	**90.20**	**81.72**	**80.19**	**73.90**	**57.42**	**50.85**	**66.70**	**54.61**	**45.01**

Removing alignment leaves only reconstruction in the second stage and reduces performance even with MCPR initialization. Thus, predictive initialization does not replace subsequent contrastive learning. Furthermore, removing verification degrades performance while retaining alignment, indicating that the treatment of cross-sample relations matters beyond enforcing consistency within samples. This finding supports using shared-view agreement to moderate potential false-negative repulsion without introducing pseudo-positive pairs. These results support the two-stage design: learning shared information for initialization and refining common representations through relation-verified contrastive alignment.

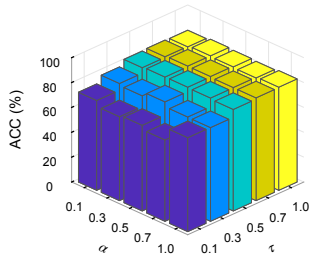
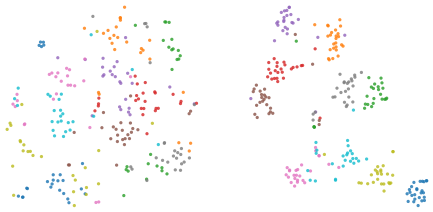


Figure 2: ACC vs. α and τ.

Parameter Analysis. We investigate the sensitivity of FIMCPA's clustering performance to the contrastive weight α and temperature τ on Caltech-5V. Both parameters are varied within the range of 0.1, 0.3, 0.5, 0.7, 1. The weight α balances contrastive learning against reconstruction. Figure 2 shows lower performance at small temperatures, such as $\tau = 0.1$. A small temperature sharpens the similarity distribution and may overemphasize highly similar negative pairs. At higher temperatures, clustering performance recovers and becomes relatively insensitive to variations in the contrastive weight α. This suggests that precise adjustment of the reconstruction–alignment balance is less critical in this region. Empirically, we set $\alpha = \tau = 0.5$ for all datasets.

Qualitative study on common representations. To examine how cross-view predictive supervision shapes common representation initialization, we visualize the global model's common representation matrix $\mathbf{H}$ immediately after pretraining. Figure 3 presents t-SNE visualizations (van der Maaten & Hinton, 2008) of the same samples from a selected multi-view client under reconstruction pretraining and MCPR pretraining.

In Figure 3(a), several classes remain scattered or intermixed, suggesting that within-view reconstruction alone does not necessarily produce common representations with clear class separation. In contrast, Figure 3(b) shows more distinct and separable features after MCPR pretraining. These observations support the benefit of cross-view predictive supervision for learning common representations that provide a better initialization for subsequent contrastive alignment.



(a) Reconstruction pretraining. (b) MCPR pretraining.

Figure 3: Visualization of pretrained common representations.

5.3 Further Analysis

Robustness to Missing Views. To examine robustness to sample-level view missingness, we vary the incomplete-sample rate r from 0.1 to 0.9 on UCI, Fashion, and CIFAR10-MV while retaining the client view configurations. Figure 4 shows that clustering performance generally declines as more samples become incomplete, reflecting the challenge of learning from reduced observations. Nevertheless, FIMCPA achieves the highest mean performance at every tested rate, maintaining its advantage even under severe missingness. These results support the effectiveness of FIMCPA in exploiting available observations to learn common representations when complete multi-view samples become scarce, extending its applicability beyond the default missingness setting.

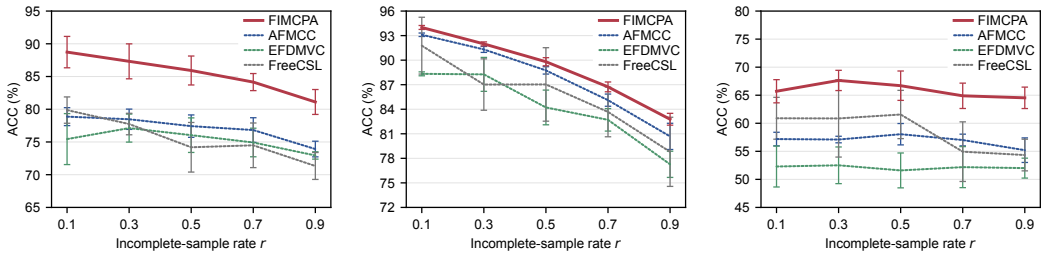


Figure 4: ACC under increasing incomplete-sample rates on (a) UCI, (b) Fashion, and (c) CIFAR10-MV. Error bars denote sample standard deviations over five paired runs; see Appendix B.5.

Sensitivity to Client Composition. We examine ratios of full-view, partial-view, and single-view clients of 2:1:1, 1:2:1, and 1:1:2 at $r = 0.5$. Full-view and partial-view clients are both multi-view clients. Table 3 shows lower mean performance at 1:2:1 than at 2:1:1, despite the same ratio of multi-view to single-view clients. Thus, the proportion of multi-view clients alone does not characterize cross-view learning conditions. Nevertheless, 1:2:1 yields higher mean performance than 1:1:2 across all four datasets with the full-view proportion unchanged. This supports the value of partial multi-view observations, which can provide target–input pairs for predictive pretraining and shared-view evidence for relation verification. These findings support FIMCPA's design for exploiting available observations beyond full-view clients.

Table 3: ACC (%) by client ratio ($r = 0.5$).

Dataset (clients)	2:1:1	1:2:1	1:1:2
UCI (8)	85.92 ± 2.21	85.66 ± 2.32	81.30 ± 0.79
CCV (16)	24.45 ± 0.75	23.27 ± 1.62	21.41 ± 0.48
Fashion (24)	89.81 ± 0.49	86.73 ± 0.66	85.12 ± 0.35
OutdoorScene (8)	73.90 ± 1.79	72.49 ± 0.67	70.93 ± 0.34

6 Conclusion

In this paper, we propose FIMCPA for federated incomplete multi-view clustering with client-level view heterogeneity and sample-level view missingness. Specifically, masked cross-view predictive pretraining is designed to learn a common representation that captures information shared across observed views. Relation-verified contrastive alignment promotes consistency between common and view-specific representations while addressing unequal sample contributions and potential false-negative repulsion. Both learning stages are supported by observation-weighted aggregation to account for differences in the available observations across clients. Theoretical analysis characterizes when shared-view agreement improves relation precision and how verification errors affect the contrastive objective. Experiments on eight benchmark configurations demonstrate that FIMCPA outperforms the compared federated and centralized baselines. Future work will investigate more heterogeneous client distributions and explicit privacy protection.

ETHICS STATEMENT

This study uses established benchmark datasets and involves no new human-subject data collection. Experiments simulate federated data distributions rather than deployment across real institutions. Keeping raw data local does not by itself provide a formal privacy guarantee.

AI USE STATEMENT

Generative AI tools assisted with literature search, experiment-design discussions, implementation and testing, mathematical formulation and proof development, result interpretation, and manuscript preparation. Reported numerical results derive from experimental runs, not AI-generated observations. The authors retain full responsibility for the accuracy and integrity of the submission.

REPRODUCIBILITY STATEMENT

Appendix A provides the theoretical assumptions and proofs. Appendices B and C describe the algorithm, training settings, evaluation protocol, and dataset configurations.

REFERENCES

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007. URL https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf.

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. In *Computer Vision – ECCV 2022*, pp. 348–367, 2022. doi: 10.1007/978-3-031-19836-6_20. URL https://doi.org/10.1007/978-3-031-19836-6_20.

Guoqing Chao, Zhenghao Zhang, Lei Meng, Jie Wen, and Dianhui Chu. Federated incomplete multi-view clustering with globally fused graph guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7417–7429. PMLR, 2025. URL https://proceedings.mlr.press/v267/chao25a.html.

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 *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3498–3506, 2023. doi: 10.1145/3581783.3612027. URL https://doi.org/10.1145/3581783.3612027.

Xinyue Chen, Yazhou Ren, Jie Xu, Fangfei Lin, Xiaorong Pu, and Yang Yang. Bridging gaps: Federated multi-view clustering in heterogeneous hybrid views. In *Advances in Neural Information Processing Systems*, volume 37, pp. 37020–37049, 2024. doi: 10.52202/079017-1168. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/412fb8623bf8b6d56fb6285ea295447e-Abstract-Conference.html.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8871–8889. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25az.html.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8765–8775, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/63c3ddcc7b23daa1e42dc41f9a44a873-Abstract.html.

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 *Advances in Neural Information Processing Systems*, volume 36, pp. 44847–44859, 2023. doi: 10.52202/075280-1943. URL https://papers.nips.cc/paper_files/paper/2023/file/8c64bc3f7796d31caa7c3e6b969bf7da-Paper-Conference.pdf.

Yuzhuo Dai, Jiaqi Jin, Zhibin Dong, Siwei Wang, Xinwang Liu, En Zhu, Xihong Yang, Xinbiao Gan, and Yu Feng. Imputation-free and alignment-free: Incomplete multi-view clustering driven by consensus semantic learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5071–5081, 2025. doi: 10.1109/CVPR52734.2025.00478. URL https://openaccess.thecvf.com/content/CVPR2025/html/Dai_Imputation-free_and_Alignment-free_Incomplete_Multi-view_Clustering_Driven_by_Consensus_Semantic_CVPR_2025_paper.html.

Xiaojian Ding, Lin Zhao, Xian Li, and Xiaoying Zhu. Incomplete multi-view clustering via hierarchical semantic alignment and cooperative completion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 58096–58116, 2025. doi: 10.52202/085713-1739. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4b2a917e30e1bb1aff055b4d8c6c081c-Paper-Conference.pdf.

Robert Duin. Multiple features. UCI Machine Learning Repository, 1998. URL https://archive.ics.uci.edu/dataset/72/multiple.

Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. Robust contrastive multi-view clustering against dual noisy correspondence. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121401–121421, 2024. doi: 10.52202/079017-3857. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/dbe81b08f7dc4dd8b43bc62dedfd9662-Abstract-Conference.html.

Dong Huang, Chang-Dong Wang, and Jian-Huang Lai. Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11388–11402, 2023. doi: 10.1109/TKDE.2023.3236698. URL https://doi.org/10.1109/TKDE.2023.3236698.

Shudong Huang, Wei Shi, Zenglin Xu, Ivor W. Tsang, and Jiancheng Lv. Efficient federated multi-view learning. *Pattern Recognition*, 131:108817, 2022. doi: 10.1016/j.patcog.2022.108817. URL https://doi.org/10.1016/j.patcog.2022.108817.

Tri Huynh, Simon Kornblith, Matthew R. Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2785–2795, 2022. URL https://openaccess.thecvf.com/content/WACV2022/html/Huynh_Boosting_Contrastive_Self-Supervised_Learning_With_False_Negative_Cancellation_WACV_2022_paper.html.

Xiaorui Jiang, Zhongyi Ma, Yulin Fu, Yong Liao, and Pengyuan Zhou. Heterogeneity-aware federated deep multi-view clustering towards diverse feature representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9184–9193, 2024. doi: 10.1145/3664647.3681302. URL https://doi.org/10.1145/3664647.3681302.

Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, pp. 1–8, 2011. doi: 10.1145/1991996.1992025. URL https://www.ee.columbia.edu/ln/dvmm/publicationPage/Publi/icmr11:consumervideo.html.

Jiaqi Jin, Siwei Wang, Zhibin Dong, Xinwang Liu, and En Zhu. Deep incomplete multi-view clustering with cross-view partial sample and prototype alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11600–11609, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Jin_Deep_Incomplete_Multi-View_Clustering_With_Cross-View_Partial_Sample_and_Prototype_CVPR_2023_paper.html.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. URL https://bottou.org/papers/lecun-98h.

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 *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 178, 2004. doi: 10.1109/CVPR.2004.383. URL https://doi.org/10.1109/CVPR.2004.383.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. COMPLETER: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11174–11183, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Lin_COMPLETER_Incomplete_Multi-View_Clustering_via_Contrastive_Prediction_CVPR_2021_paper.html.

Suyuan Liu, Hao Yu, Hao Tan, Ke Liang, Siwei Wang, Shengju Yu, En Zhu, and Xinwang Liu. Efficient federated incomplete multi-view clustering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40100–40114. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25cq.html.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982. doi: 10.1109/tit.1982.1056489. URL https://doi.org/10.1109/tit.1982.1056489.

Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(13):14193–14201, 2024. doi: 10.1609/aaai.v38i13.29330. URL https://doi.org/10.1609/aaai.v38i13.29330.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Jingyu Pu, Chenhang Cui, Xinyue Chen, Yazhou Ren, Xiaorong Pu, Zhifeng Hao, Philip S. Yu, and Lifang He. Adaptive feature imputation with latent graph for deep incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14633–14641, 2024. doi: 10.1609/aaai.v38i13.29380. URL https://ojs.aaai.org/index.php/AAAI/article/view/29380.

Yazhou Ren, Xinyue Chen, Jie Xu, Jingyu Pu, Yonghao Huang, Xiaorong Pu, Ce Zhu, Xiaofeng Zhu, Zhifeng Hao, and Lifang He. A novel federated multi-view clustering method for unaligned and incomplete data fusion. *Information Fusion*, 108:102357, 2024. doi: 10.1016/j.inffus.2024.102357. URL https://doi.org/10.1016/j.inffus.2024.102357.

Huayi Tang and Yong Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21090–21110. PMLR, 2022. URL https://proceedings.mlr.press/v162/tang22c.html.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL https://www.jmlr.org/papers/v9/vandermaaten08a.html.

Jiazhen Wang, Xinyue Chen, Shuaiyu Liu, Zican He, Yazhou Ren, Yi Wang, and Shuyin Xia. Bridging inter-view and client heterogeneity: Federated multi-view clustering under non-IID data. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 4949–4957, 2026. doi: 10.24963/ijcai.2026/551. URL https://www.ijcai.org/proceedings/2026/551.

Bingjun Wei, Xuemei Cao, Jiafen Liu, Haoyang Liang, and Xin Yang. Enhanced federated deep multi-view clustering under uncertainty scenario. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(32):26806–26813, 2026. doi: 10.1609/aaai.v40i32.39891. URL https://ojs.aaai.org/index.php/AAAI/article/view/39891.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL https://arxiv.org/abs/1708.07747.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16051–16060, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Xu_Multi-Level_Feature_Learning_for_Contrastive_Multi-View_Clustering_CVPR_2022_paper.html.

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 *Advances in Neural Information Processing Systems*, volume 36, pp. 1119–1131. Curran Associates, Inc., 2023. doi: 10.52202/075280-0054. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/03b13b0db740b95cb741e007178ef5e5-Abstract-Conference.html.

Zhenhao Xu, Huazhu Chen, Weiwei Wang, and Wenhui Duan. Asynchronous federated multimodal constrained clustering under arbitrary modality missingness. *Neurocomputing*, 698: 134329, 2026. doi: 10.1016/j.neucom.2026.134329. URL https://doi.org/10.1016/j.neucom.2026.134329.

Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1055–1069, 2023. doi: 10.1109/TPAMI.2022.3155499. URL https://doi.org/10.1109/TPAMI.2022.3155499.

A Theory and Proofs

The analysis addresses three questions: which observation statistics client roles fail to determine; when shared-view agreement improves relation precision and how verification errors perturb the loss; and which parameter barycenter the aggregation rule computes. The loss and gradient results are conditional on fixed representations or candidates, with additional assumptions stated below. They do not establish a clustering benefit from MCPR or end-to-end training convergence.

A.1 Proof of client-mask insufficiency

Let $p_{cuv} = \sum_{i \in \mathcal{I}_c} R_{ciu} R_{civ}$ count samples observed in both views u and v.

Proposition 1 (Client-mask insufficiency)**.** *If client c owns at least two views and two samples, $(A_{c:}, N_c)$ does not determine all $\{n_{cv}\}_v$, $\{p_{cuv}\}_{u,v}$, or $\{\mathcal{S}_{cij}\}_{i,j}$ for every admissible within-client mask.*

Fix two acquired views $u, v \in \mathcal{V}_c$ and distinct samples $i, j \in \mathcal{I}_c$. Construct one admissible mask M that observes u on every sample and observes v on i and j. Construct M' identically except that $M'_{civ} = 0$. (Other owned entries may be held fixed.) Both masks have the same $(A_{c:}, N_c)$ and every sample retains view u. Nevertheless, n_{cv} and p_{cuv} each decrease by one under M', while $v \in \mathcal{S}_{cij}$ under M and $v \notin \mathcal{S}_{cij}$ under M'. A deterministic function of the shared input $(A_{c:}, N_c)$ must return the same output for both masks and therefore cannot recover these realized quantities in every admissible case. This proves Proposition 1.

A.2 Support-based calibration and spherical consensus

For unit vectors $\mathbf{b}, \mathbf{u}$ and $\rho \in [0, 1]$, consider the spherical consensus problem underlying equation 3:

$$\min_{\|\mathbf{h}\|_2 = 1} \frac{1}{2} \|\mathbf{h} - \mathbf{b}\|_2^2 + \frac{\rho}{2} \|\mathbf{h} - \mathbf{u}\|_2^2. \tag{13}$$

Proposition 2 (Uniqueness of support-only calibration)**.** *If $w : 2^{[V]} \to [0, 1]$ is view-permutation invariant, additive on disjoint supports, and satisfies $w(\varnothing) = 0, w([V]) = 1$, then $w(S) = |S|/V$.*

Permutation invariance makes every singleton mass equal: write $w(\{v\}) = a$ for all v. Repeated disjoint additivity gives $w(S) = \sum_{v \in S} w(\{v\}) = |S|a$. Since $w([V]) = Va = 1$, we obtain $a = 1/V$ and therefore $w(S) = |S|/V$, proving Proposition 2.

For unit vectors $\mathbf{h}, \mathbf{b}, \mathbf{u}$, expanding equation 13 leaves a constant minus $\mathbf{h}^\top(\mathbf{b}+\rho\mathbf{u})$. If $\mathbf{b}+\rho\mathbf{u} \neq \mathbf{0}$, Cauchy–Schwarz shows that its unique unit-norm minimizer is $\mathbf{h} = (\mathbf{b} + \rho\mathbf{u})/\|\mathbf{b} + \rho\mathbf{u}\|_2$. This solution coincides with the clamped update only when $\|\mathbf{b} + \rho\mathbf{u}\|_2 \geq \epsilon$ and the two branch outputs are unit vectors. For a nonzero resultant below ϵ, the clamped output has norm below one and is not a feasible solution of the spherical problem. At a zero resultant, every unit vector minimizes the objective, whereas the implementation returns zero. Complete support sets $\rho = 1$ and recovers the uncalibrated sum under the same numerical normalization convention.

A.3 Properties and implementation of relation verification

Complete the neighborhood definition by setting $\mathcal{N}_{cv}(i) = \varnothing$ when $i \notin \mathcal{B}_{cv}$ or $\nu_{cv} < 2$.

Proposition 3 (Conservative relation verification)**.** *Fix the minibatch, binary realized support, model state, and any random or tie-breaking choices. Assume unobserved inputs are replaced by zero before encoding. Then G_c is symmetric with zero diagonal, and its removal sets are disjoint from the positive sets. Changing an input with $R_{civ} = 0$ cannot affect G_c or $\mathcal{L}_{\mathrm{C}}^{\mathrm{c}}$. If no pair shares two views, verification removes no nodes. A removal excludes the pair's direct denominator term without adding a cross-sample positive term.*

Mutual membership gives symmetry, and the explicit convention $G_{cii} = 0$ gives a zero diagonal, including singleton minibatches. The condition $j \neq i$ in equation 7 keeps removals disjoint from same-sample positives; the observed-view nodes, common representations, and neighborhoods use only observations with $R = 1$ (with unobserved inputs replaced by zero). Thus an incorrect removal may suppress useful repulsion but cannot turn the pair into a pseudo-positive. The remaining denominator probabilities are renormalized, so their gradients need not remain unchanged.

The implementation first appends one common node for every minibatch row and then appends view nodes selected by $R_{civ} = 1$ in registered-view order. If id_α denotes the sample ID of node α, the sample-ID matrix used for positives is $P_{\alpha\beta} = \mathbf{1}[\mathrm{id}_\alpha = \mathrm{id}_\beta]\mathbf{1}[\alpha \neq \beta]$, with zero diagonal, encoding the same-sample positive relation in equation 26. The implementation averages positive losses per node, scatters node losses back to their sample IDs, and divides by the number of nodes of each sample before averaging rows. This implements Equation 9 for variable support sizes.

For every view, the code selects the observed row indices, computes detached cosine similarities, removes self-similarity, and takes the number of neighbors in equation 5. Intersecting directed neighborhoods with their transpose gives mutual membership. A pair is marked as verified only when its agreement count equals its number of shared observed views and that number is at least two, as specified in equation 6. Indexing this sample-level matrix by the semantic-node sample IDs excludes all node pairs belonging to distinct samples whose relation has been verified. The positive-pair matrix uses equality of sample IDs, so verified relations cannot become positives.

A.4 Maximality under shared-view agreement constraints

For a shared view define $C^v_{cij} = \mathbf{1}[j \in \mathcal{N}_{cv}(i)]\mathbf{1}[i \in \mathcal{N}_{cv}(j)]$. Formally, $\mathfrak{G}_c$ is the family of binary symmetric, zero-diagonal masks $\widetilde{G}_c$ satisfying

$$\begin{aligned} &\widetilde{G}_{cij} = 0 \quad \text{if } |\mathcal{S}_{cij}| < 2, \\ &\widetilde{G}_{cij} \leq C^v_{cij} \quad \forall v \in \mathcal{S}_{cij}. \end{aligned} \tag{14}$$

Proposition 4 (Maximal admissible verification mask)**.** *For fixed support and neighborhoods, G_c is the unique pointwise maximum of $\mathfrak{G}_c$ and maximizes $\sum_{i<j} w_{ij}\widetilde{G}_{cij}$ for every $w_{ij} \geq 0$.*

For any $\widetilde{G}_c \in \mathfrak{G}_c$, a pair with fewer than two shared views has $\widetilde{G}_{cij} = G_{cij} = 0$. For every other pair, equation 14 gives $\widetilde{G}_{cij} \leq C^v_{cij}$ for each $v \in \mathcal{S}_{cij}$. Since all variables are binary,

$$\widetilde{G}_{cij} \leq \prod_{v \in \mathcal{S}_{cij}} C^v_{cij} = G_{cij}.$$

Conversely, G_c is binary, symmetric, zero-diagonal, obeys the two-view eligibility condition, and is bounded by every factor in its product; hence $G_c \in \mathfrak{G}_c$. It is therefore the unique pointwise maximum. After multiplication by $w_{ij} \geq 0$ and summation, it maximizes every stated weighted objective. If every pair selected by G_c has positive weight, any different admissible mask omits at least one such pair and has strictly smaller weight.

For a fixed anchor node, take w_{ij} to be the sum of its nonnegative exponentiated similarity scores over nodes belonging to sample j (and analogously for every anchor). Monotonicity then proves the direct removed-score statement. The positive set remains the same-sample set in equation 26, so maximal cancellation creates no cross-sample positive. This proves Proposition 4.

A.5 Neighborhood size under class imbalance

Let π_{cvk} be the empirical proportion of class k among the $\nu_{cv} \geq 2$ observed samples. For a query of class k, exactly $\nu_{cv}\pi_{cvk} - 1$ other observed samples have that class. The triangle inequality gives

$$|q_{cv} - (\nu_{cv}\pi_{cvk} - 1)| \leq \left|q_{cv} - \left(\frac{\nu_{cv}}{K} - 1\right)\right| + \nu_{cv}\left|\pi_{cvk} - \frac{1}{K}\right|. \tag{15}$$

This bound compares the chosen capacity with the number of available same-class candidates; it makes no independence assumption about sampling. The labels defining π_{cvk} are used only in this analysis, not in neighborhood selection. A balanced capacity can under-cover majority clusters and over-cover minority clusters. Shared-view mutuality does not remove this mismatch or guarantee that selected neighbors belong to the query's class. Even under exact balance, $\nu_{cv} = Km$ with $K \geq 2$ gives $q_{cv} = m$, whereas only $m - 1$ other samples share the query's class. Thus directed neighborhoods need not be class-pure, even before mutuality and cross-view intersection are applied.

A.6 Dependence-corrected cross-view precision

Let $Z_{ij} = 1$ mean that samples i, j belong to the same unknown cluster and $B^v_{ij} = 1$ denote mutual nearest-neighbor agreement. Condition on the realized-support event $H_{\mathcal{S}} = \{\mathcal{S}_{cij} = \mathcal{S}\}$ and set $A_{\mathcal{S}} = \cap_{v\in\mathcal{S}}\{B^v_{ij} = 1\}$, $\pi_{\mathcal{S}} = \Pr(Z_{ij} = 1 \mid H_{\mathcal{S}})$, $\mathrm{TPR}_{v,\mathcal{S}} = \Pr(B^v_{ij} = 1 \mid Z_{ij} = 1, H_{\mathcal{S}})$, and $\mathrm{FPR}_{v,\mathcal{S}} = \Pr(B^v_{ij} = 1 \mid Z_{ij} = 0, H_{\mathcal{S}})$. Define $\kappa_{z,\mathcal{S}} := \Pr(A_{\mathcal{S}} \mid Z_{ij} = z, H_{\mathcal{S}}) / \prod_{v\in\mathcal{S}} \Pr(B^v_{ij} = 1 \mid Z_{ij} = z, H_{\mathcal{S}})$. All probabilities below are taken over a fixed sampling distribution of distinct sample pairs and minibatch realizations within a fixed client, with the model parameters and verifier rule held fixed.

Proposition 5 (Dependence-corrected cross-view precision)**.** *Assume $|\mathcal{S}| \geq 2$, $\Pr(H_{\mathcal{S}}) > 0$, and $0 < \pi_{\mathcal{S}} < 1$. If the conditional joint and marginal agreement probabilities defining both dependence factors are positive, then*

$$\frac{\Pr(Z_{ij} = 1 \mid A_{\mathcal{S}}, H_{\mathcal{S}})}{\Pr(Z_{ij} = 0 \mid A_{\mathcal{S}}, H_{\mathcal{S}})} = \frac{\pi_{\mathcal{S}}}{1 - \pi_{\mathcal{S}}} \frac{\kappa_{1,\mathcal{S}}}{\kappa_{0,\mathcal{S}}} \prod_{v\in\mathcal{S}} \frac{\mathrm{TPR}_{v,\mathcal{S}}}{\mathrm{FPR}_{v,\mathcal{S}}}. \tag{16}$$

On $H_{\mathcal{S}}$, $G_{cij} = 1$ is equivalent to $A_{\mathcal{S}}$. The posterior odds exceed the support-conditional prior odds exactly when the dependence-adjusted product of likelihood ratios exceeds one.

Fix a realized shared-view set $\mathcal{S}$ and condition throughout on $H_{\mathcal{S}}$. The definitions of $\kappa_{1,\mathcal{S}}, \kappa_{0,\mathcal{S}}$ and Bayes' rule give

$$\frac{\Pr(Z_{ij} = 1 \mid A_{\mathcal{S}}, H_{\mathcal{S}})}{\Pr(Z_{ij} = 0 \mid A_{\mathcal{S}}, H_{\mathcal{S}})} = \frac{\pi_{\mathcal{S}}}{1 - \pi_{\mathcal{S}}} \frac{\Pr(A_{\mathcal{S}} \mid Z_{ij} = 1, H_{\mathcal{S}})}{\Pr(A_{\mathcal{S}} \mid Z_{ij} = 0, H_{\mathcal{S}})} = \frac{\pi_{\mathcal{S}}}{1 - \pi_{\mathcal{S}}} \frac{\kappa_{1,\mathcal{S}}}{\kappa_{0,\mathcal{S}}} \prod_{v\in\mathcal{S}} \frac{\mathrm{TPR}_{v,\mathcal{S}}}{\mathrm{FPR}_{v,\mathcal{S}}}.$$

Normalizing the displayed odds yields the stated posterior precision; comparison with the $H_{\mathcal{S}}$-conditional prior odds yields the condition. Conditional independence given $(Z_{ij}, H_{\mathcal{S}})$ makes both dependence factors one. This proves Proposition 5. The dependence factors account for departures from conditional independence across views.

A.7 Verified-denominator risk decomposition

Fix an anchor and let D^0 be its self-excluded denominator before relation verification. Let $\mathcal{E}$ contain the cross-sample nodes removed by verification, and let $\mathcal{T}$ contain the cross-sample nodes with

the anchor's unknown class. Their intersection and differences give correct removals $\mathcal{C} = \mathcal{E} \cap \mathcal{T}$, incorrect removals $\mathcal{F} = \mathcal{E} \setminus \mathcal{T}$, and missed false negatives $\mathcal{M} = \mathcal{T} \setminus \mathcal{E}$. Normalize their score masses by D^0 as $\gamma_{\mathrm{C}}, \gamma_{\mathrm{F}}, \gamma_{\mathrm{M}}$.

The exact loss difference between relation verification and oracle verification is

$$\ell^V - \ell^T = \log \frac{1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}}}{1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{M}}}. \tag{17}$$

Proposition 6 (Verified-denominator risk decomposition)**.** *Assume all relevant denominators are positive. Let ℓ^0, ℓ^V, ℓ^C, and ℓ^T remove cross-sample node sets $\varnothing$, $\mathcal{E}$, $\mathcal{C}$, and $\mathcal{T}$, respectively, with the same positive set and positive logits. In addition to equation 17,*

$$\begin{aligned} \ell^V - \ell^0 &= \log(1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}}), \\ |\ell^V - \ell^T| &\le \frac{\gamma_{\mathrm{F}} + \gamma_{\mathrm{M}}}{\min\{1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}},\, 1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{M}}\}}, \\ 0 \le \ell^C - \ell^V &\le \frac{\gamma_{\mathrm{F}}}{1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}}}. \end{aligned} \tag{18}$$

Let $W(\mathcal{A})$ denote the sum of exponentiated similarity scores over a node set $\mathcal{A}$. By construction, $\mathcal{C}$, $\mathcal{F}$, and $\mathcal{M}$ are disjoint where required, and

$$\begin{aligned} D^V &= D^0 - W(\mathcal{C}) - W(\mathcal{F}) = D^0(1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}}), \\ D^T &= D^0 - W(\mathcal{C}) - W(\mathcal{M}) = D^0(1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{M}}). \end{aligned}$$

For each removal rule $X \in \{0, V, C, T\}$, let D^X denote its denominator and write $\bar{t}_{\mathcal{P}} = |\mathcal{P}|^{-1} \sum_{p \in \mathcal{P}} t_p$ for the mean positive logit. The multi-positive loss has the identity $\ell^X = \log D^X - \bar{t}_{\mathcal{P}}$, so all cross-sample removals leave the positive term unchanged. Subtracting two negative log-softmax losses therefore cancels this common term and gives the displayed log-ratio identities.

For positive x, y, the mean-value theorem gives $|\log x - \log y| \le |x - y| / \min\{x, y\}$. Taking $x = 1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}}$ and $y = 1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{M}}$ proves the oracle-gap bound. Finally,

$$\ell^C - \ell^V = \log \frac{1 - \gamma_{\mathrm{C}}}{1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}}} \le \frac{\gamma_{\mathrm{F}}}{1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}}},$$

where the last inequality is $-\log(1 - u) \le u/(1 - u)$ for $u \in [0, 1)$. Nonnegativity follows because removing $\mathcal{F}$ can only decrease the denominator. This proves Proposition 6. The proof is algebraic and makes no independence or view-quality assumption; it bounds contrastive-loss perturbation, not downstream clustering error.

A.8 Proof of oracle-gradient and conditional assignment stability

Let $\mathbf{p}^V, \mathbf{p}^T$ be the softmax distributions over denominator nodes after relation verification and oracle removals, respectively, each extended by zero to the union of their supports. Define $\delta = \min\{1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}}, 1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{M}}\}$, and write $\gamma_{\mathrm{O}} = 1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}} - \gamma_{\mathrm{M}}$. Here $\mathbf{y}$ denotes the anchor representation before the update. The analysis differentiates a single anchor loss with respect to this representation. Candidate vectors $\mathbf{y}_j$ are held fixed after normalization, together with the positive and removal sets. Thus the numerical logit is $t_j(\mathbf{y}) = \mathrm{norm}(\mathbf{y})^\top \mathbf{y}_j / \tau$; the already normalized candidates are not normalized again in this expression.

Theorem 1 (Oracle-gradient and conditional assignment stability)**.** *Assume fixed candidate vectors with $\|\mathbf{y}_j\|_2 \le 1$, positive denominators, $\tau > 0$, and $\|\mathbf{y}\|_2 > \epsilon$, with $0 < \epsilon < 1$. For the cosine loss in equation 8,*

$$\|\nabla_{\mathbf{y}} \ell^V - \nabla_{\mathbf{y}} \ell^T\|_2 \le \frac{\|\mathbf{p}^V - \mathbf{p}^T\|_1}{\tau \|\mathbf{y}\|_2} \le \frac{2(\gamma_{\mathrm{F}} + \gamma_{\mathrm{M}})}{\tau \|\mathbf{y}\|_2 \delta}. \tag{19}$$

Take one representation-space step of size $\eta \ge 0$ from the same anchor under each loss and apply $\mathrm{norm}(\cdot)$ to both results. If the two pre-normalization norms are at least $\gamma > 0$, the resulting representations differ by at most $2\eta(\gamma_{\mathrm{F}} + \gamma_{\mathrm{M}})/(\tau \|\mathbf{y}\|_2 \gamma \delta)$. For fixed Lloyd centers with $K \ge 2$, let the oracle-step assignment have margin $\Delta > 0$ and maximum competing-center distance $D > 0$. The relation-verified step preserves this assignment whenever

$$\frac{2\eta(\gamma_{\mathrm{F}} + \gamma_{\mathrm{M}})}{\tau \|\mathbf{y}\|_2 \gamma \delta} < \frac{\Delta}{2D}. \tag{20}$$

Specifically, let $\mathbf{y}_T$ be the oracle-step representation after $\mathrm{norm}(\cdot)$ and k^* its unique nearest center. The squared-distance margin is $\Delta = \min_{k \neq k^*}(\|\mathbf{y}_T - \boldsymbol{\mu}_k\|_2^2 - \|\mathbf{y}_T - \boldsymbol{\mu}_{k^*}\|_2^2)$, and $D = \max_{k \neq k^*} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k^*}\|_2$.

Let the self-excluded softmax distribution be $\mathbf{p}^0$. The disjoint sets $\mathcal{C}, \mathcal{F}, \mathcal{M}$ have $\mathbf{p}^0$ masses $\gamma_{\mathrm{C}}, \gamma_{\mathrm{F}}, \gamma_{\mathrm{M}}$, and their complement has mass $\gamma_{\mathrm{O}} = 1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}} - \gamma_{\mathrm{M}}$. The relation-verified distribution retains the complement and $\mathcal{M}$, with normalizer $r_V = 1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{F}} = \gamma_{\mathrm{O}} + \gamma_{\mathrm{M}}$; the oracle distribution retains the complement and $\mathcal{F}$, with normalizer $r_T = 1 - \gamma_{\mathrm{C}} - \gamma_{\mathrm{M}} = \gamma_{\mathrm{O}} + \gamma_{\mathrm{F}}$. Summing the absolute probability difference separately on $\mathcal{M}$, $\mathcal{F}$, and their common retained complement gives

$$\|\mathbf{p}^V - \mathbf{p}^T\|_1 = \frac{\gamma_{\mathrm{M}}}{r_V} + \frac{\gamma_{\mathrm{F}}}{r_T} + \frac{\gamma_{\mathrm{O}}|\gamma_{\mathrm{M}} - \gamma_{\mathrm{F}}|}{r_V r_T}. \tag{21}$$

Writing $\delta = \min\{r_V, r_T\}$, the first two terms are at most $(\gamma_{\mathrm{M}} + \gamma_{\mathrm{F}})/\delta$. The remaining factor satisfies $\gamma_{\mathrm{O}}/(r_V r_T) \leq 1/\max\{r_V, r_T\} \leq 1/\delta$ and $|\gamma_{\mathrm{M}} - \gamma_{\mathrm{F}}| \leq \gamma_{\mathrm{M}} + \gamma_{\mathrm{F}}$. Together these give $\|\mathbf{p}^V - \mathbf{p}^T\|_1 \leq 2(\gamma_{\mathrm{F}} + \gamma_{\mathrm{M}})/\delta$.

Set $\widehat{\mathbf{y}} = \mathbf{y}/\|\mathbf{y}\|_2$. Since $\|\mathbf{y}\|_2 > \epsilon$, the cosine derivative uses the Jacobian $(I - \widehat{\mathbf{y}}\widehat{\mathbf{y}}^\top)/\|\mathbf{y}\|_2$. The common positive-numerator gradient cancels, and the chain rule gives

$$\nabla_{\mathbf{y}} \ell^V - \nabla_{\mathbf{y}} \ell^T = \frac{I - \widehat{\mathbf{y}}\widehat{\mathbf{y}}^\top}{\tau \|\mathbf{y}\|_2} \sum_j (p_j^V - p_j^T)\mathbf{y}_j. \tag{22}$$

The projection has operator norm at most one. The triangle inequality and $\|\mathbf{y}_j\|_2 \leq 1$ therefore give equation 19. If $\bar{\mathbf{y}}^X = \mathbf{y} - \eta \nabla_{\mathbf{y}} \ell^X$ for $X \in \{V, T\}$, then $\|\bar{\mathbf{y}}^V - \bar{\mathbf{y}}^T\|_2$ is at most η times that gradient bound. For vectors with norms $r = \|\mathbf{a}\|_2 \geq \gamma > 0$ and $s = \|\mathbf{b}\|_2 \geq \gamma$, the identity

$$\|\mathbf{a} - \mathbf{b}\|_2^2 = (r - s)^2 + rs \left\| \frac{\mathbf{a}}{r} - \frac{\mathbf{b}}{s} \right\|_2^2$$

gives a $1/\gamma$ Lipschitz bound for ordinary normalization. If $\gamma \geq \epsilon$, both vectors lie outside or on the clamping boundary, so the same bound holds for $\mathrm{norm}(\cdot)$. If $0 < \gamma < \epsilon$, $\mathrm{norm}(\cdot)$ equals Euclidean projection onto the closed ball of radius ϵ, followed by division by ϵ. This projection is nonexpansive, giving a $1/\epsilon \leq 1/\gamma$ bound. Thus, in both cases,

$$\| \mathrm{norm}(\mathbf{a}) - \mathrm{norm}(\mathbf{b}) \|_2 \leq \frac{\|\mathbf{a} - \mathbf{b}\|_2}{\gamma}.$$

Applying this bound to the two updates gives

$$\left\| \mathrm{norm}\left(\bar{\mathbf{y}}^V\right) - \mathrm{norm}\left(\bar{\mathbf{y}}^T\right) \right\|_2 \leq \frac{2\eta(\gamma_{\mathrm{F}} + \gamma_{\mathrm{M}})}{\tau \|\mathbf{y}\|_2 \gamma \delta}. \tag{23}$$

This step does not require $\gamma \geq \epsilon$ or unit outputs. Finally, expanding the difference of squared distances cancels the quadratic representation term. For $d_k(\mathbf{x}) = \|\mathbf{x} - \boldsymbol{\mu}_k\|_2^2 - \|\mathbf{x} - \boldsymbol{\mu}_{k^*}\|_2^2$, this gives $d_k(\mathbf{x}) - d_k(\mathbf{z}) = 2(\mathbf{x} - \mathbf{z})^\top(\boldsymbol{\mu}_{k^*} - \boldsymbol{\mu}_k)$. Each competing-center gap therefore changes by at most $2D$ times the representation perturbation. The strict condition in Equation 20 therefore keeps every positive oracle-step margin positive. This proves Theorem 1. For unit anchors, the factor $\|\mathbf{y}\|_2$ equals one. The result concerns one anchor loss and fixed candidates and centers; it does not bound the joint network update, candidate-role gradients from other anchors, or later changes to neighborhoods and centers.

A.9 Observation-weighted parameter barycenter

In equation 11, $\boldsymbol{\vartheta}_c$ denotes any local parameter block. For the view-specific encoder–decoder $\boldsymbol{\omega}_c^v = (\phi_{c,v}, \theta_{c,v})$ and the input-view block $\mathbf{W}_{c,v}$ of the first fusion matrix $\mathbf{W}_c = [\mathbf{W}_{c,1}|\cdots|\mathbf{W}_{c,V}]$, the weight is $q_c = n_{cv}$. For the remaining shared fusion parameters $\boldsymbol{\psi}_c^{\mathrm{sh}}$, it is $q_c = m_c$. If a view has zero total support, the implementation copies its block from the first client rather than evaluating a zero-denominator average. This convention is separate from the positive-support barycenter result below.

For a view-specific block, let $Q_v(\theta) = \sum_c n_{cv}\|\theta - \theta_c^v\|_2^2$. When $\sum_c n_{cv} > 0$, its Hessian is the positive scalar $2\sum_c n_{cv}$ times the identity and its unique stationary point is

$$\theta = \frac{\sum_c n_{cv}\theta_c^v}{\sum_c n_{cv}},$$

which is the view-specific instance of equation 11. Terms with $n_{cv} = 0$ vanish from both the objective and update, formalizing exclusion of unsupported view blocks. The shared-block and first-fusion-layer rules apply the identical argument to total realized mass and to each input-view block, respectively. These weights count observed support, not nonzero gradient contributions to every block. In particular, single-view clients have $m_c > 0$ but provide no predictive gradient to the fusion network during MCPR. The result characterizes the chosen parameter-distance objective; it does not show that observation-count weights minimize the training risk.

B Implementation and Reproducibility

Support and local-data notation. Let $\mathcal{V}_c = \{v : A_{cv} = 1\}$ and $\mathcal{C}_F = \{c : |\mathcal{V}_c| = V\}$, $\mathcal{C}_P = \{c : 1 < |\mathcal{V}_c| < V\}$, and $\mathcal{C}_S = \{c : |\mathcal{V}_c| = 1\}$, indexed by f, p, s, respectively. The local dataset is $\mathcal{D}_c = \{\{\mathbf{x}_i^v : R_{civ} = 1\} : i \in \mathcal{I}_c\}$; $p_{cuv} = \sum_{i\in\mathcal{I}_c} R_{ciu}R_{civ}$ counts co-observed samples. Other support quantities are defined in Section 3.

Common representation and local objectives. Let $\Theta = ((\phi_v, \theta_v)_{v=1}^V, \psi)$ collect the model parameters, with Θ_c denoting client c's local copy. In implementation, each view-specific reconstruction and prediction term in Eqs. (2) and (4) is additionally normalized by its feature dimension. The contrastive weight $\alpha = 0.5$ is applied relative to the resulting dimension-normalized reconstruction loss. Numerically, $\mathrm{norm}(\mathbf{a}) = \mathbf{a}/\max(\|\mathbf{a}\|_2, \epsilon)$ with $0 < \epsilon < 1$; an absent slot is zero, not a synthesized latent. The output has unit norm only when the input norm is at least ϵ.

MCPR uses the eligible-target loss in equation 4. Using these implementation-normalized losses, its minibatch objective is

$$\mathcal{L}_{\mathrm{pre}}^c = \begin{cases} \frac{1}{2}(\mathcal{L}_{\mathrm{R}}^c + \mathcal{L}_{\mathrm{P}}^c), & \sum_{v\in\mathcal{V}_c} |\mathcal{E}_{cv}| > 0 \\ \mathcal{L}_{\mathrm{R}}^c, & \sum_{v\in\mathcal{V}_c} |\mathcal{E}_{cv}| = 0 \end{cases}. \tag{24}$$

The branch depends on target availability, not the prediction-loss value. All eligible targets are evaluated before one backward pass. Target removal applies to the fusion slots, observed-view sum, and support fraction; contrastive training restores the original observations.

Semantic-node sets. For contrastive training the complete node set is

$$\mathcal{U}_c = \{(i, 0) : i \in \mathcal{B}_c\} \cup \{(i, v) : i \in \mathcal{B}_c, R_{civ} = 1\}. \tag{25}$$

The positive set of anchor (i, a) is

$$\mathcal{P}_{ci}^a = \{(i, b) \in \mathcal{U}_c : b \neq a\}. \tag{26}$$

Similarity is evaluated as $\mathrm{norm}(\mathbf{a})^\top \mathrm{norm}(\mathbf{b})$, which also defines the value at zero and clamps small norms. View-specific latent representations are normalized once for the similarity matrix; the common representation is normalized again at this step.

Clustering objective and initialization. With the learned representations fixed, federated clustering applies Lloyd updates (Lloyd, 1982) to the objective

$$\min_{\{\boldsymbol{\mu}_k\}_{k=1}^K} \sum_c \sum_{i\in\mathcal{I}_c} \min_{k\in[K]} \|\mathbf{h}_i - \boldsymbol{\mu}_k\|_2^2. \tag{27}$$

For initialization, client c runs local K_c-means with $K_c = \min(K, N_c)$. Write $\mathbf{r}_{ck}$ for its returned centers and b_{ci} for its final assignments. The uploaded weight of each center is

$$w_{ck} = \sum_{i\in\mathcal{I}_c} \mathbf{1}[b_{ci} = k], \qquad k \in [K_c]. \tag{28}$$

Algorithm 1 FIMCPA: predictive pretraining, contrastive alignment, and federated clustering

Require: Client data $\{\mathbf{x}_i^v, R_{civ}\}_{c,i,v}$, number of clusters K, initial parameters Θ^{init}, pretraining phases P, contrastive-training rounds T, and local epoch budgets (Appendix B.2)
All aggregations use equation 11 with the zero-support convention in Appendix A.9.
Stage 1: Masked cross-view predictive pretraining
Initialize the server parameters $\Theta \leftarrow \Theta^{\text{init}}$
for $t = 1, \ldots, P$ **do**
  Server broadcasts Θ
  for all clients c in parallel **do**
    Initialize local parameters $\Theta_c \leftarrow \Theta$
    for all local minibatches over the prescribed epochs for this phase **do**
      Compute $\mathcal{L}_{\text{R}}^c$ from the original observations
      Compute $\mathcal{L}_{\text{P}}^c$ by summing eligible-target losses as in equation 4; hold out each target separately
      Take one gradient update of Θ_c using equation 24; use reconstruction alone when $\sum_{v \in \mathcal{V}_c} |\mathcal{E}_{cv}| = 0$
    end for
    Upload the updated Θ_c and support counts (n_{cv}, m_c)
  end for
  Update Θ by aggregating the uploaded local parameter blocks
end for
Set $\Theta^0 \leftarrow \Theta$; use the original observation masks for contrastive training
Stage 2: Relation-verified contrastive alignment
for $t = 1, \ldots, T$ **do**
  Server broadcasts Θ^{t-1}
  for all clients c in parallel **do**
    Initialize local parameters $\Theta_c \leftarrow \Theta^{t-1}$
    for all local minibatches over the prescribed local epochs **do**
      Encode observed views and compute $\mathbf{h}_i$ using equation 3
      Form within-sample positive cliques and detached mutual neighborhoods using equation 5
      Compute G_{cij} by equation 6, requiring at least two shared views and agreement in every shared view
      Take one gradient update of Θ_c using equation 10 and the sample-balanced loss in equation 9; exclude verified cross-sample pairs only from the denominator
    end for
    Upload the updated Θ_c and support counts (n_{cv}, m_c)
  end for
  Obtain Θ^t by aggregating the uploaded local parameter blocks
end for
Federated clustering
Server broadcasts the final aggregated model Θ^T
for all clients c in parallel **do**
  Encode $\{\mathbf{h}_i : i \in \mathcal{I}_c\}$; run local K_c-means with $K_c = \min(K, N_c)$; upload the returned centers and counts from equation 28
end for
Server runs weighted K-means on representatives to initialize $\{\boldsymbol{\mu}_k\}_{k=1}^K$
repeat
  Server broadcasts centers; clients assign samples by equation 29 and upload only $\{\mathbf{S}_{ck}, n_{ck}\}_{k=1}^K$
  The server updates centers using equation 12 and retains the previous center for an empty global cluster
until the Frobenius center shift is at most 10^{-5} or 20 updates have been completed
Server broadcasts the final centers; clients recompute their assignments
return local assignments induced by the final global cluster centers

The implementation uploads all K_c pairs, including any zero-count centers. Local fitting uses K-means++ initialization (Arthur & Vassilvitskii, 2007) and 10 restarts, with the run seed plus the zero-based client index as its random seed. The server fits weighted K-means to these pairs with 100 restarts and the run seed. Returned local centers are not recomputed from the final assignments and need not equal the means of their assigned samples if fitting stops before the assignments converge. During refinement, each client assigns its samples to the broadcast centers by

$$\widehat{k}_{ci} = \arg\min_{k \in [K]} \|\mathbf{h}_i - \boldsymbol{\mu}_k\|_2^2. \tag{29}$$

For each cluster, it sends the sum $\mathbf{S}_{ck} = \sum_{i:\widehat{k}_{ci}=k} \mathbf{h}_i$ and count $n_{ck} = |\{i \in \mathcal{I}_c : \widehat{k}_{ci} = k\}|$. The server applies equation 12 to nonempty global clusters and retains the previous center for an empty one, as specified in Algorithm 1.

B.1 Exact-arithmetic Lloyd equivalence and numerical implementation

First work in exact real arithmetic on the same fixed representations. Assume identical initial centers, identical deterministic tie handling, and identical empty-cluster rules. At any iteration, each sample sees the same centers in both procedures and therefore receives the same assignment. Additivity gives

$$\begin{aligned} \sum_c \mathbf{S}_{ck} &= \sum_{c,i:\widehat{k}_{ci}=k} \mathbf{h}_i, \\ \sum_c n_{ck} &= \left|\{(c,i) : \widehat{k}_{ci} = k\}\right|. \end{aligned} \tag{30}$$

Equation 12 is consequently the arithmetic mean used by centralized Euclidean Lloyd. Induction proves algebraic equality of every later center and assignment at matching iteration indices. Identical stopping rules and budgets then give identical final outputs.

The numerical implementation computes distances using 32-bit floating-point (FP32) features and centers; ties are resolved by choosing the lowest center index. Local sums accumulate in 64-bit floating-point (FP64) format before sums and counts are serialized as FP32. The server stacks these arrays in client order, reduces them in FP64, divides nonempty sums by counts, and casts centers to FP32; empty clusters retain their previous centers. FP32 counts are exact at the evaluated client sizes, which are below 2^{24}. Refinement stops after at most 20 updates or when the Frobenius norm of the center change is at most 10^{-5}. Clients then reassign using the final centers.

The exact-arithmetic proof does not establish bitwise equivalence to an arbitrary centralized implementation. Such a comparison requires matching the complete numerical execution, including distances, accumulation order, casts, tie handling, and stopping rules. Aggregate statistics also provide no privacy guarantee; secure aggregation would be an additional deployment mechanism.

B.2 Training configurations

FIMCPA uses the same predictive and contrastive objectives on every dataset, with $\alpha = 0.5$ and $\tau = 0.5$. Batch size is 256, the learning rate is 3×10^{-4}, the latent and common dimensions are both 20, and pretraining is divided into 20 aggregation phases.

Table 4 specifies the shared architecture. All layers are fully connected with biases, ReLU hidden activations, and linear outputs; no batch normalization or dropout is used. Fusion follows equation 3.

Table 4: Network architecture of FIMCPA.

Module	Layer widths
View-v encoder	$D_v \to 500 \to 500 \to 2000 \to 20$
View-v decoder	$20 \to 2000 \to 500 \to 500 \to D_v$
Fusion network	$20V \to 256 \to 20$

Both stages use Adam (Kingma & Ba, 2015) with $(\beta_1, \beta_2) = (0.9, 0.999)$, stabilizer 10^{-8}, constant learning rate, and zero weight decay, without AMSGrad or gradient clipping. Moments reset at each pretraining phase or contrastive round and persist across its local epochs.

Table 5 lists the training configurations used for FIMCPA in the main comparison.

Baselines retain their method-native training procedures and use the same paired seeds $10, 20, 30, 40, 50$.

Table 5: FIMCPA training configurations used for the main comparison.

Dataset	Clients	Pretraining epochs	Local epochs	Rounds
Caltech-5V	8	600	9	15
UCI	8	400	10	5
HW	8	100	10	20
OutdoorScene	8	400	25	10
CCV	16	500	10	60
Fashion	24	200	10	50
CIFAR10-MV	40	150	3	100
MNIST-3V	40	80	5	25

B.3 Evaluation protocol

We report means and sample standard deviations over five paired seeds $10, 20, 30, 40, 50$. FIMCPA uses final global cluster centers; FMCSC and AFMCC refine shared initial centers from the first client using client-wise K-means. Predictions are concatenated across clients before evaluation, without client-wise label matching or score averaging. ACC uses one global Hungarian matching; NMI and ARI use the same concatenated predictions. Centralized baselines are evaluated on pooled samples. EFDMVC supports full-, partial-, and single-view clients.

Observation-mask interfaces. Table 6 summarizes how baselines use $R = A \odot M$. Centralized methods pool client data and the corresponding masks; federated methods retain client-specific masks. Zero filling alone does not make an absent entry an observed training target.

Table 6: Observation-mask interfaces of baseline implementations.

Method	Use of sample-level observation masks
DSIMVC	Pooled R identifies observed entries for scaling and incomplete-data training. Donor search is expanded when needed without lowering the imputation threshold; only eligible observations supply donors, not missing targets. Clustering averages prediction probabilities over originally observed views.
FreeCSL	Per-view columns of pooled R enter reconstruction, contrastive, and graph-learning routines. Common features used for center estimation and final K-means average only the observed views of each sample.
HSACC	Pooled R selects all-view-complete rows for the released training procedure. At inference, the mask distinguishes observed views from those handled by cooperative latent inference; clustering uses concatenated latent representations.
FMCSC	Reconstruction and view-conditioned contrastive terms use observed rows; absent latent slots are masked before concatenation-based fusion. Cross-pattern transfer requires all target-pattern views to be observed and the reference sample to retain additional views. Pattern-based aggregation and client-wise K-means are retained.
AFMCC	Each subset-training step uses only samples observing every selected view and is skipped when none qualify. Reconstruction and contrastive terms use the eligible rows; mean fusion excludes absent views. Personalized aggregation and client-wise K-means are retained.
EFDMVC	Reconstruction uses observed entries, contrastive terms use the corresponding observed or co-observed samples, and missing-view latent slots are masked before fusion.
H2-FedMVC	Reconstruction and view-conditioned contrastive terms use observed rows. Fusion slots are masked after encoding by R and the expert's structural view mask.

B.4 Computation and communication

MCPR adds one prediction pass per eligible target view, using the existing encoders, fusion network, and decoder. Target losses share one backward pass, so their activations accumulate in memory.

Contrastive training constructs a dense semantic-node similarity matrix; verification adds detached within-view similarities, neighbor selection, and pair masks. Neither module adds trainable parameters or message types, although both incur local computation and storage. For C participating clients and P_Θ FP32 model parameters, a full-model upload and download cost $8CP_\Theta$ bytes per exchange, excluding support counts and transport overhead. Fixed observation counts can be cached. Clustering exchanges local centers and counts, followed by cluster sums, counts, and updated centers (Algorithm 1). Its total cost depends on the number of refinement iterations and repeated evaluations; aggregate communication is not necessarily smaller than a one-time embedding upload.

B.5 Missing-View Evaluation Protocol

We compare FIMCPA, AFMCC, EFDMVC, and FreeCSL on UCI, Fashion, and CIFAR10-MV at $r \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, using five paired seeds $10, 20, 30, 40, 50$. Client counts are 8, 24, and 40, respectively, with a full/partial/single ratio of 2:1:1. Client partitions and acquisition masks remain fixed across rates; sample masks follow Appendix C and are shared across methods. Method-specific scaling uses observed entries only; absent entries are zero-filled and masked. For each dataset, all four methods keep their respective training configurations fixed across incomplete-sample rates. Table 7 reports ACC; $r = 0.5$ reuses the main-comparison results.

Table 7: ACC under incomplete-sample rates on UCI, Fashion, and CIFAR10-MV. Values report the mean ± sample standard deviation (%) over five paired runs; the $r = 0.5$ entries reuse the main-table results.

Dataset	r	FIMCPA	AFMCC	EFDMVC	FreeCSL
UCI	0.1	88.73±2.40	78.87±1.37	75.44±3.89	79.87±2.01
UCI	0.3	87.32±2.67	78.47±1.55	77.12±2.14	77.74±1.64
UCI	0.5	85.92±2.21	77.42±1.71	76.04±2.64	74.19±3.79
UCI	0.7	84.15±1.30	76.82±1.88	74.92±2.18	74.49±3.41
UCI	0.9	81.11±1.90	73.92±1.20	72.94±0.56	71.33±2.05
Fashion	0.1	94.00±0.24	93.10±0.21	88.33±0.24	91.77±3.47
Fashion	0.3	91.99±0.24	91.31±0.36	88.26±2.07	87.02±3.14
Fashion	0.5	89.81±0.49	88.75±0.45	84.22±2.12	87.03±4.49
Fashion	0.7	86.73±0.60	85.11±0.75	82.70±1.37	83.67±3.04
Fashion	0.9	82.78±0.72	80.68±1.60	77.27±1.60	78.79±4.23
CIFAR10-MV	0.1	65.71±2.07	57.18±1.21	52.29±3.65	60.89±3.73
CIFAR10-MV	0.3	67.63±1.80	57.10±0.58	52.51±3.27	60.86±6.89
CIFAR10-MV	0.5	66.70±2.62	58.05±1.91	51.60±3.11	61.56±4.30
CIFAR10-MV	0.7	64.89±2.26	57.01±1.04	52.18±3.66	54.94±5.31
CIFAR10-MV	0.9	64.54±1.91	55.20±2.19	52.00±1.78	54.34±2.82

B.6 Ablation Design and Results

All variants retain observation-weighted aggregation and share client partitions, observation masks, five paired seeds, initial weights, and training schedules up to the evaluated model. Equal schedules do not imply equal computation. The full/partial/single client ratio is 2:1:1; Full results reuse Table 1.

Variant definitions. *w/o prediction* replaces MCPR with $\mathcal{L}_{\mathrm{R}}^{\mathrm{c}}$ alone, retaining reconstruction pretraining and its schedule. *w/o target masking* retains the eligible-target sets and prediction objective, but substitutes $\mathbf{h}_i$ for $\mathbf{h}_i^{(-v)}$ in equation 4, leaving the target view available to the predictor. *w/o alignment* retains MCPR but uses reconstruction-only updates during subsequent training. *w/o verification* retains all cross-sample denominator terms while preserving same-sample positives and their weights. Table 8 reports means and sample standard deviations over five paired runs. These interventions assess clustering performance, not verifier precision or interactions between jointly removed components.

Table 8: Complete ablation results on four datasets.

Dataset	Variant	ACC	NMI	ARI
UCI	w/o prediction	77.25 ± 5.75	67.72 ± 4.11	61.04 ± 5.74
	w/o target masking	81.86 ± 3.71	74.55 ± 2.13	69.06 ± 3.51
	w/o alignment	84.19 ± 2.42	74.55 ± 2.46	70.09 ± 3.67
	w/o verification	85.61 ± 2.25	76.10 ± 2.01	72.46 ± 3.32
	FIMCPA (Full)	$\mathbf{85.92 \pm 2.21}$	$\mathbf{76.58 \pm 1.98}$	$\mathbf{72.96 \pm 3.32}$
HW	w/o prediction	89.32 ± 1.62	80.36 ± 1.96	78.47 ± 2.61
	w/o target masking	89.05 ± 3.00	81.20 ± 2.14	79.30 ± 3.19
	w/o alignment	67.28 ± 3.26	59.37 ± 2.07	50.80 ± 2.85
	w/o verification	88.34 ± 3.13	80.02 ± 2.63	77.84 ± 3.76
	FIMCPA (Full)	$\mathbf{90.20 \pm 2.12}$	$\mathbf{81.72 \pm 2.71}$	$\mathbf{80.19 \pm 3.52}$
OutdoorScene	w/o prediction	72.78 ± 1.71	56.23 ± 1.71	49.65 ± 2.19
	w/o target masking	72.65 ± 2.01	56.12 ± 1.86	49.21 ± 2.73
	w/o alignment	63.26 ± 4.44	46.41 ± 3.64	38.66 ± 3.79
	w/o verification	73.10 ± 2.35	56.29 ± 2.13	49.58 ± 3.18
	FIMCPA (Full)	$\mathbf{73.90 \pm 1.79}$	$\mathbf{57.42 \pm 1.78}$	$\mathbf{50.85 \pm 2.45}$
CIFAR10-MV	w/o prediction	65.763 ± 3.372	54.295 ± 0.684	43.135 ± 3.057
	w/o target masking	66.303 ± 2.767	54.325 ± 1.471	44.222 ± 2.219
	w/o alignment	56.156 ± 5.080	46.762 ± 2.391	34.613 ± 3.808
	w/o verification	60.704 ± 1.970	49.245 ± 1.254	38.258 ± 2.004
	FIMCPA (Full)	$\mathbf{66.698 \pm 2.618}$	$\mathbf{54.609 \pm 0.897}$	$\mathbf{45.013 \pm 1.401}$

B.7 Client-Composition Evaluation Protocol

We fix $r = 0.5$ and vary the full/partial/single client ratio over 2:1:1, 1:2:1, and 1:1:2. UCI, CCV, Fashion, and OutdoorScene use 8, 16, 24, and 8 clients, respectively. Within each dataset and paired seed, client sample partitions, model configurations, training budgets, objectives, aggregation, and clustering remain fixed across ratios. Observation masks follow Appendix C; changing acquisition support can also change the total observed information. We use paired scenario and optimization seeds $10, 20, 30, 40, 50$ and the shared hyperparameters in Appendix B.2. OutdoorScene uses 400 pretraining epochs and 10 contrastive-training rounds with 25 local epochs per round across ratios. Table 9 reports all metrics (mean and sample SD); the 2:1:1 results reuse Table 1.

Table 9: Client-composition results (mean $\pm$ sample SD, %; five paired runs). Ratios follow full/partial/single order.

Dataset	Ratio	ACC	NMI	ARI
UCI	2:1:1	85.920 ± 2.211	76.579 ± 1.977	72.961 ± 3.316
	1:2:1	85.660 ± 2.319	75.906 ± 1.780	72.380 ± 3.484
	1:1:2	81.300 ± 0.787	72.059 ± 0.726	66.336 ± 1.071
CCV	2:1:1	24.453 ± 0.745	18.779 ± 0.176	8.765 ± 0.298
	1:2:1	23.269 ± 1.619	17.798 ± 0.421	8.347 ± 0.606
	1:1:2	21.411 ± 0.475	15.977 ± 0.454	7.359 ± 0.473
Fashion	2:1:1	89.808 ± 0.494	82.326 ± 0.654	79.783 ± 0.892
	1:2:1	86.732 ± 0.658	78.447 ± 0.701	74.773 ± 0.778
	1:1:2	85.116 ± 0.352	76.237 ± 0.321	71.673 ± 0.433
OutdoorScene	2:1:1	73.899 ± 1.793	57.416 ± 1.778	50.846 ± 2.451
	1:2:1	72.493 ± 0.670	55.573 ± 1.283	48.490 ± 1.022
	1:1:2	70.930 ± 0.343	53.385 ± 1.048	46.245 ± 0.650

Multi-view availability for individual samples does not establish pair-wise verification eligibility. These composition comparisons do not isolate component effects or compare methods under changing compositions.

C Data and Missing-View Protocols

The primary comparison uses class-stratified, approximately equal-size client partitions with both client-level view heterogeneity and sample-level view missingness. All methods share the same client partitions and observation masks. At incomplete-sample rate r, each client with $k \geq 2$ acquired views selects $\lfloor rN_c \rfloor$ samples uniformly without replacement to be incomplete; the main comparison uses $r = 0.5$. Each selected sample independently retains u views, with u uniform on $\{1, \ldots, k-1\}$ and a uniformly chosen size-u subset of the acquired views. Other samples retain all acquired views; single-view clients retain their sole view. Appendix B.5 reports the completed UCI, Fashion, and CIFAR10-MV sweep using this same generation rule at other incomplete-sample rates. Labels are used to form the class-stratified partitions and to compute external clustering metrics. For FIMCPA, they do not enter the training objectives or federated inference.

Client-level view allocation. The main comparison uses a full/partial/single client ratio of 2:1:1. Full-view clients acquire all V views. For partial-view clients, view-set sizes cycle through $2, \ldots, V-1$ and are shuffled using the scenario seed. For each size, the allocator enumerates candidate subsets and selects one by lexicographically minimizing four quantities after its addition: the range and sum of squared per-view ownership counts, followed by the range and sum of squared view-pair co-ownership counts. These counts are accumulated over partial-view clients only; a seeded random candidate order breaks ties. Single-view clients cycle through a seeded permutation of the V views. The complete list of full-, partial-, and single-view supports is then shuffled and assigned to the client partitions. This allocation uses neither feature values nor class labels; it is distinct from the class-stratified sample partitioning described above.

View indices follow the input-view order in Table 10.

C.1 Dataset configurations and sources

Table 10 lists benchmark sizes and input-view dimensions. The descriptions below distinguish original collections from the multi-view configurations used here.

Table 10: Multi-view benchmark configurations. Feature dimensions follow our input-view order.

Dataset	Samples	Classes	Views	Feature dimensions
Caltech-5V	1,400	7	5	40 / 254 / 928 / 512 / 1,984
UCI	2,000	10	3	240 / 76 / 6
HW	2,000	10	6	216 / 76 / 64 / 6 / 240 / 47
OutdoorScene	2,688	8	4	512 / 432 / 256 / 48
CCV	6,773	20	3	20 / 20 / 20
Fashion	10,000	10	3	784 / 784 / 784
CIFAR10-MV	50,000	10	3	512 / 2,048 / 1,024
MNIST-3V	60,000	10	3	342 / 1,024 / 64

Caltech-5V. Caltech-5V is a seven-class subset of Caltech-101 (Li Fei-Fei et al., 2004), matching the sample count and feature dimensions reported in ESFMC (Chen et al., 2025).

UCI and HW. Both use the Multiple Features handwritten-numeral collection (Duin, 1998): UCI uses three views, while HW uses all six, as in ESFMC (Chen et al., 2025). They are different view configurations of the same source collection.

OutdoorScene. Its sample count, class count, and feature dimensions match the Out-Scene benchmark reported in FastMICE (Huang et al., 2023).

CCV. CCV originates from the Columbia Consumer Video database (Jiang et al., 2011). We use the pre-extracted `CCV.mat` features with three 20-dimensional views, a configuration also reported in EFIMVC (Liu et al., 2025), rather than extracting the original audio-visual features.

Fashion. Fashion is derived from Fashion-MNIST (Xiao et al., 2017), with the same benchmark dimensions as Multi-Fashion in FMCSC (Chen et al., 2024). Its construction groups different images of the same category into one multi-view sample (Cui et al., 2023).

CIFAR10-MV. CIFAR10-MV uses the 50,000-image CIFAR-10 training split (Krizhevsky, 2009), represented by three feature views with dimensions $512/2048/1024$.

MNIST-3V. MNIST-3V uses the pre-extracted `MNIST_fea_Per1.mat` features derived from MNIST (LeCun et al., 1998), with dimensions $342/1024/64$, also reported in EFIMVC (Liu et al., 2025). We use its aligned feature arrays and generate observation masks under Appendix C, rather than using the file's embedded missingness mask.

Input preprocessing. CCV, MNIST-3V, and CIFAR10-MV features are min–max scaled to $[0, 1]$ separately for each feature dimension within each view. All compared methods receive the same input features, client partitions, and observation masks within each benchmark.

D Notation Summary

We write $[n] = \{1, \ldots, n\}$ and $\mathbf{1}[\cdot]$ for indicators; c indexes any client role.

Table 11: Principal notation. Semantic-slot indices a, b and local dummy indices are scoped to their defining equations.

Symbol	Meaning
$[C], c$; $[V], v, u$; $[K], k$	Client, registered-view, and cluster index sets and indices.
$\mathcal{C}_F, f$; $\mathcal{C}_P, p$; $\mathcal{C}_S, s$	Full-view, partial-view, and single-view client sets and their role-specific indices; generic formulas use c for their union.
$\mathcal{D}_c, \mathcal{I}_c, N_c, i$	Private local dataset, its non-overlapping sample-index set, its size, and a sample index.
A_{cv}, M_{civ}, R_{civ}	Structural acquisition, conditional observation, and realized support, with $R_{civ} = A_{cv}M_{civ}$.
$\mathcal{V}_c$	Structurally available views of client c; its cardinality defines the full-view, partial-view, or single-view acquisition role.
$\mathbf{x}_i^v, \mathbf{z}_i^v, \mathbf{h}_i$	Observed input, view-specific latent representation, and common representation.
n_{cv}, p_{cuv}, m_c	Observed sample count for one view, co-observed sample count for a view pair, and total observed sample–view count at client c.
$\mathcal{B}_c, \mathcal{B}_{cv}, \mathcal{U}_c$	Local minibatch, its samples observed in view v, and its semantic-node set.
$\mathcal{E}_{cv}, \mathbf{h}_i^{(-v)}, \rho_i^{(-v)}$	Samples eligible to predict observed target view v, common representation with that target held out, and its reduced realized-support fraction.
$\mathcal{L}_{\mathrm{P}}^c, \mathcal{L}_{\mathrm{pre}}^c$	Observed-target predictive loss and MCPR pretraining objective, with the fallback in equation 24 when no eligible prediction targets are available.
$\mathcal{P}_{ci}^a$	Same-sample positive nodes for semantic-node slot a.
$\mathcal{A}_{ci}^a$	Admissible denominator candidates for anchor (i, a): all non-self nodes except nodes of other samples verified as related to the anchor.
$q_{cv}, \mathcal{N}_{cv}(i), G_{cij}$	Neighborhood size, directed nearest-neighbor candidate set, and binary relation-verification mask requiring mutual agreement in all shared views.
$\mathcal{L}_{\mathrm{R}}^c, \mathcal{L}_{\mathrm{C}}^c, \mathcal{L}^c$	Observed reconstruction, relation-verified contrastive loss, and contrastive-training objective, with $\mathcal{L}^c = \mathcal{L}_{\mathrm{R}}^c + \alpha\mathcal{L}_{\mathrm{C}}^c$.
$\boldsymbol{\omega}^v, \mathbf{W}_v, \boldsymbol{\psi}^{\mathrm{sh}}$	View autoencoder block, corresponding fusion-input columns, and the remaining shared fusion parameters.
$\boldsymbol{\vartheta}_c, q_c$	Generic local parameter block and its aggregation weight: n_{cv} for view-specific blocks or m_c for shared fusion parameters.
$\boldsymbol{\mu}_k, \widehat{k}_{ci}, \mathbf{S}_{ck}, n_{ck}$	Global cluster center, induced assignment, and client sum/count sufficient statistics for federated Lloyd refinement.
t, P, T	Communication-round index, number of pretraining phases, and number of contrastive-training rounds. In experiments, r denotes the incomplete-sample rate.

D_v and d denote input and representation dimensions; α, τ, and ϵ denote the contrastive weight, temperature, and normalization clamp. Representation and shared-view quantities follow Sections 3 and 4.2.

E COMPLETE CLUSTERING RESULTS

Table 12 supplements Table 1 with sample standard deviations ($\mathrm{ddof} = 1$) over the same five runs, with means and standard deviations rounded to two decimal places; ranking markers are retained from the original results.

Table 12: Complete clustering performance (mean ± sample SD, %; five paired runs under the prescribed training configurations). DSIMVC, FreeCSL, and HSACC are centralized. Best and second-best means are **bold** and underlined.

Method	Caltech-5V			UCI		
	ACC	NMI	ARI	ACC	NMI	ARI
DSIMVC	45.76±1.83	36.40±4.08	26.10±4.97	62.82±5.86	61.18±2.58	46.87±4.00
FreeCSL	79.36±4.16	66.61±2.79	62.40±4.33	74.19±3.79	71.59±3.19	62.25±4.87
HSACC	40.96±4.73	41.64±7.49	19.34±8.89	66.89±6.73	70.53±3.83	51.41±10.75
FMCSC	71.26±3.21	53.15±3.18	49.47±3.52	75.56±2.34	66.88±1.97	60.44±2.45
AFMCC	71.49±7.51	55.92±6.12	51.45±7.94	77.42±1.71	70.75±1.06	63.52±1.59
EFDMVC	65.67±1.75	52.04±2.37	44.77±2.00	76.04±2.64	71.24±2.03	63.12±3.31
H2-FedMVC	78.66±3.44	60.39±4.68	59.28±5.41	83.75±1.95	73.00±2.11	68.97±2.92
FIMCPA	**79.51±3.21**	**66.94±2.80**	**63.65±3.68**	**85.92±2.21**	**76.58±1.98**	**72.96±3.32**
Method	HW			OutdoorScene		
	ACC	NMI	ARI	ACC	NMI	ARI
DSIMVC	57.62±4.42	54.19±3.69	40.07±4.81	49.21±1.84	39.92±3.65	27.94±2.21
FreeCSL	82.57±2.15	77.70±1.31	71.98±2.60	53.45±3.92	46.85±1.86	33.54±2.58
HSACC	65.81±4.47	69.32±4.10	47.33±11.05	46.97±4.27	45.24±5.43	28.15±8.61
FMCSC	77.29±1.95	68.82±1.64	63.38±1.79	60.36±4.60	43.94±3.90	36.49±4.21
AFMCC	76.83±1.11	68.30±0.96	61.41±1.60	68.39±2.79	50.27±2.72	43.43±3.08
EFDMVC	80.53±4.82	73.86±2.16	68.90±3.24	61.38±2.07	48.20±1.19	37.88±1.17
H2-FedMVC	84.71±1.99	71.86±2.83	69.57±3.49	70.98±1.51	53.54±1.80	47.01±2.07
FIMCPA	**90.20±2.12**	**81.72±2.71**	**80.19±3.52**	**73.90±1.79**	**57.42±1.78**	**50.85±2.45**
Method	CCV			Fashion		
	ACC	NMI	ARI	ACC	NMI	ARI
DSIMVC	17.50±0.42	14.76±0.38	5.55±0.29	74.08±2.51	72.18±1.01	62.78±1.95
FreeCSL	20.74±1.10	17.01±1.00	7.61±0.69	87.03±4.49	81.27±1.27	77.53±3.79
HSACC	14.91±0.86	11.00±0.46	3.11±0.45	81.07±5.87	78.10±1.21	70.70±4.79
FMCSC	18.10±0.76	13.49±0.41	5.38±0.09	83.44±1.97	73.00±2.03	69.62±2.49
AFMCC	19.70±0.80	15.95±0.77	6.56±0.42	88.75±0.45	80.80±0.53	78.23±0.71
EFDMVC	20.56±0.74	16.95±0.87	7.25±0.50	84.22±2.12	76.75±1.35	71.38±2.48
H2-FedMVC	16.62±1.51	12.67±1.10	4.80±0.66	89.56±0.53	81.91±0.78	79.72±0.95
FIMCPA	**24.45±0.75**	**18.78±0.18**	**8.77±0.30**	**89.81±0.49**	**82.33±0.65**	**79.78±0.89**
Method	CIFAR10-MV			MNIST-3V		
	ACC	NMI	ARI	ACC	NMI	ARI
DSIMVC	50.43±0.97	43.89±1.16	29.51±0.86	96.43±3.91	93.47±3.40	93.63±5.37
FreeCSL	61.56±4.30	53.93±2.71	42.55±4.73	98.66±0.05	96.07±0.10	97.03±0.11
HSACC	37.09±2.49	34.54±2.17	9.01±1.89	61.99±5.33	70.68±6.06	37.29±10.28
FMCSC	49.20±3.30	37.85±1.71	27.95±2.65	93.18±2.61	85.65±2.92	85.77±4.60
AFMCC	58.05±1.91	46.61±0.89	37.84±1.22	98.49±0.03	95.65±0.08	96.68±0.06
EFDMVC	51.60±3.11	42.86±0.97	31.09±1.23	98.27±0.07	95.12±0.15	96.22±0.15
H2-FedMVC	61.58±3.94	46.10±4.38	38.09±4.51	98.39±0.32	95.33±0.72	96.48±0.71
FIMCPA	**66.70±2.62**	**54.61±0.90**	**45.01±1.40**	**99.00±0.06**	**96.93±0.14**	**97.78±0.13**

F ADDITIONAL RELATED WORK

Alignment under incomplete observations. CPSPAN uses pair-observed data to learn cross-view instance correspondence and aligns cluster prototypes to calibrate incomplete distributions (Jin et al., 2023). In FIMCPA, predictive initialization uses held-out observed views, while subsequent relation verification screens cross-sample negatives using shared observed views within each client.

Graph-assisted feature imputation. AGDIMC combines partial latent graphs with adaptive imputation of missing embedded features using cross-view soft cluster assignments and global cluster centroids (Pu et al., 2024). In FIMCPA, prediction targets are restricted to observed views, and relation verification uses available representations without generating genuinely missing view features.