TCRIMVC: TARGET-CONDITIONED RETRIEVAL FOR RECOVERY AND REFINEMENT IN INCOMPLETE MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Incomplete multi-view clustering (IMVC) aims to discover coherent cluster structures from multi-view data with partially missing observations, yet existing approaches often rely on dedicated cross-view predictors or generative models to infer unavailable representations, thereby introducing additional modeling assumptions and exposing downstream clustering to potential recovery errors. We revisit this problem from a retrieval perspective: relevant semantics may already exist among observed representations and can be reused as evidence for recovery. To this end, we propose a target-conditioned retrieval framework for incomplete multi-view clustering. Given the observed views of an instance and the identity of a requested view, TCRIMVC retrieves relevant latent evidence for representation construction. Missing-view representations are constructed by aggregating retrieved evidence, whereas observed representations are refined through a residual update that retains their original features. Both pathways retrieve from encoded observations, without reusing completed outputs as evidence. The resulting representations are jointly optimized through view reconstruction, global-to-view alignment, and confidence-aware balanced cross-view consistency, reducing the influence of uncertain recoveries on consensus learning. Extensive experiments on diverse benchmarks demonstrate the effectiveness and robustness of TCRIMVC under varying incomplete conditions. Further analyses validate the contributions of its retrieval, refinement, and consensus mechanisms, while cross-view image completion results illustrate the broader applicability of the proposed retrieval principle.

1 INTRODUCTION

Multi-view clustering (MVC) groups unlabeled instances by integrating several descriptions of the same underlying objects (Zhao et al., 2023; Wang, 2026; Wu et al., 2025; Yan et al., 2024). Its standard premise—that every instance is present in every view—is routinely violated by sensor failure, unavailable tests, and asynchronous data collection (Zhuang et al., 2025; Wang et al., 2026b; Yan et al., 2023a). In Incomplete Multi-View Clustering (IMVC), the missingness pattern is itself heterogeneous: two instances can expose different subsets of views, so neither direct sample-wise alignment nor ordinary feature fusion has a complete set of inputs. The central problem is therefore more specific than learning a common space. For every absent instance–view pair, a model must decide which observed information is relevant, how that information should be transferred, and how uncertain recovery should influence the downstream cluster structure.

Existing deep IMVC approaches make different compromises along this path. Predictive and generative methods learn cross-view mappings, latent-variable models, normalizing flows, or diffusion processes to synthesize unavailable content (Lin et al., 2021; 2022; Palumbo et al., 2024; Gao & Pu, 2025; Zhang et al., 2025; Jin et al., 2025; Ding et al., 2026). Prototype and graph-based methods instead recover a missing representation from cluster anchors or propagated neighborhood structure (Li et al., 2023; Jin et al., 2023; Chao et al., 2024; Du et al., 2026; Zhang et al., 2026; Liu et al., 2024). These methods can explicitly restore unavailable information, but the inferred representation is mediated by a dataset-level mapping, a generative model, a prototype, or a preconstructed graph. Imputation-free methods avoid potentially misleading completion and learn from the available sub-

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.

sets through complementarity mining, distributional aggregation, geometric alignment, or semantic consensus (Xu et al., 2022; 2023; 2024; Dai et al., 2025; Chen et al., 2026; He et al., 2026; Zhou et al., 2026); however, they do not explicitly populate the missing target-view slot. This contrast motivates a middle ground: recover a representation explicitly, but obtain it from observed evidence rather than from a separate representation generator.

Neighbor-based IMVC already provides an important starting point. DSIMVC locates semantic neighbors in the observed source views and averages their available target-view counterparts (Tang & Liu, 2022); MICA transfers the k-nearest-neighbor topology of a selected reliable view for multi-level imputation (Wang et al., 2025); and SCVT first selects a source view through an optimal-transport view graph and then aggregates corresponding target-view neighbors (Dong et al., 2025). Meanwhile, cross-sample aggregation has been used to enrich a complete-view consensus or an incomplete shared representation (Yan et al., 2023b; Wang et al., 2026a). These studies establish that observed instances can be useful recovery context, and consequently rule out any broad claim that neighbor-based completion or cross-sample attention is new. They also expose a sharper question: can evidence access be learned for each *instance–target-view request*, jointly from all views available to that instance, rather than being determined by a source-wise neighborhood, a single selected view, a global graph, or a shared-level attention block?

To address this question, we formulate incomplete multi-view representation learning as target-conditioned direct-context retrieval. Inspired by the retrieve-before-infer principle of retrieval-augmented generation (Lewis et al., 2020), TCRIMVC directly exploits observed representations as retrieval evidence rather than introducing an additional generator for missing views. Specifically, all views are encoded into a normalized shared latent space. Given a missing target view, TCRIMVC constructs a target-conditioned query by integrating the available-view semantics with the target-view identity, retrieves a sparse Top-K set of relevant observed representations, and recovers the missing representation through weighted evidence aggregation. Beyond missing-view recovery, the same mechanism refines observed representations by retrieving complementary same-view evidence from other instances for residual enhancement. The resulting representations are fused for clustering and jointly optimized with view-specific reconstruction, global-view alignment, and reliability-aware cross-view cluster consistency. In particular, retrieval confidence adaptively suppresses unreliable completions during consensus learning. In this way, TCRIMVC unifies missing-view recovery and observed-view enhancement through direct retrieval from observed evidence, without requiring a separate representation generator. The main contributions of this paper are summarized as follows:

- We introduce a new perspective on incomplete multi-view clustering (IMVC), formulating missing-view recovery as target-conditioned direct-context retrieval rather than representation generation. A query combining the available views and target-view identity selects observed latent evidence to construct the missing representation without a separate representation generator.
- We develop a unified retrieval mechanism that jointly handles missing and observed views within the same framework. Missing representations are completed through sparse aggregation of cross-view evidence, while observed representations are refined via self-excluding, same-view residual retrieval, enabling completion and representation enhancement to reinforce each other.
- We propose reliability-aware consensus learning that explicitly accounts for uncertainty in retrieved evidence. Retrieval-derived confidence adaptively modulates balanced cross-view cluster consistency, suppressing unreliable completions while preserving informative cross-view supervision for robust clustering.

2 Related Work

Completion-based IMVC. Existing methods recover missing views using cross-view predictors or generative models (Lin et al., 2021; 2022; Palumbo et al., 2024; Gao & Pu, 2025; Zhang et al., 2025; Jin et al., 2025; Ding et al., 2026), while others exploit prototypes, anchors, or graph structures (Li et al., 2023; Jin et al., 2023; Chao et al., 2024; Liu et al., 2024; Du et al., 2026; Zhang et al., 2026). More closely related methods transfer information from neighboring samples across views (Tang & Liu, 2022; Wang et al., 2025; Dong et al., 2025) or employ cross-sample aggregation to enhance shared representations (Yan et al., 2023b; Wang et al., 2026a). TCRIMVC differs

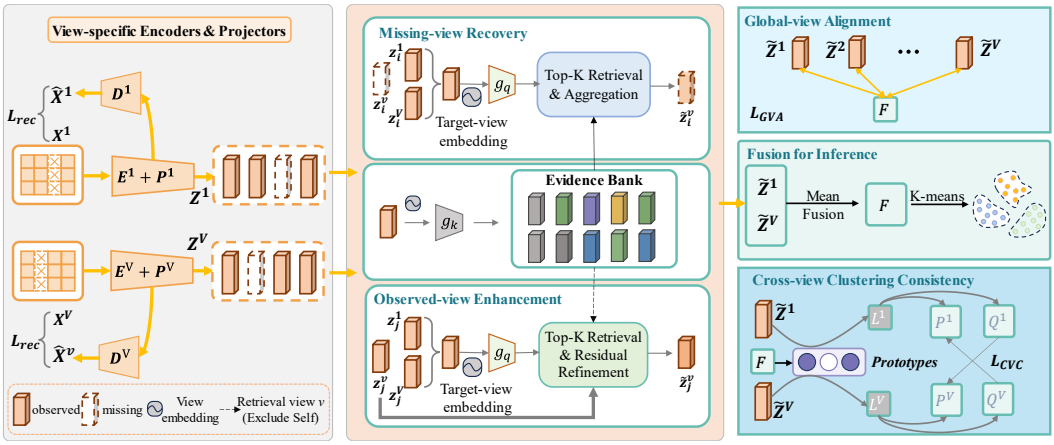


Figure 1: **Overview of TCRIMVC.** Missing and observed view slots are processed using separate recovery and enhancement evidence banks, respectively. The completed representations are aligned with the fused feature and optimized by reliability-weighted cross-view clustering consistency.

by constructing an instance–target-view query from all available views and directly retrieving observed latent evidence for both missing-view recovery and observed-view enhancement, without an additional representation generator.

Imputation-free IMVC. Imputation-free methods avoid explicitly reconstructing missing views and instead learn from available observations through complementarity mining, distribution alignment, variational aggregation, semantic consensus, or graph-based learning (Xu et al., 2022; 2023; 2024; Dai et al., 2025; Chen et al., 2026; He et al., 2026; Zhou et al., 2026). Although these methods reduce the potential influence of inaccurate imputation, they do not explicitly recover the missing target-view representations. TCRIMVC instead completes the latent view slots using traceable retrieved evidence and employs retrieval confidence to reduce the influence of unreliable completions during cross-view consensus learning.

3 Method

3.1 Problem Setup and View-Specific Encoding

Let $\mathcal{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{V}$ denote a V-view dataset with N aligned instances, where $\mathbf{X}^{(v)} \in \mathbb{R}^{N \times d_v}$. We use a binary mask $\mathbf{M} \in \{0,1\}^{N \times V}$ to characterize view availability: $m_i^{(v)} = 1$ if view v of instance i is observed and $m_i^{(v)} = 0$ otherwise. Accordingly, $\mathcal{O}_i = \{v \mid m_i^{(v)} = 1\}$ denotes the set of observed views of instance i. Following the standard transductive IMVC setting, we assume $|\mathcal{O}_i| \geq 1$ for every instance and that the number of clusters C is known.

Each view v is associated with a view-specific encoder E_v, decoder D_v, and projection head P_v. For an observed view, we compute

$$\mathbf{h}_i^{(v)} = E_v\left(\mathbf{x}_i^{(v)}\right), \qquad \widehat{\mathbf{x}}_i^{(v)} = D_v\left(\mathbf{h}_i^{(v)}\right), \qquad \mathbf{z}_i^{(v)} = \nu\left(P_v\left(\mathbf{h}_i^{(v)}\right)\right), \quad v \in \mathcal{O}_i, \tag{1}$$

where $\mathbf{z}_i^{(v)} \in \mathbb{R}^d$ lies in a shared latent space and $\nu(\mathbf{a}) = \mathbf{a}/\max(\|\mathbf{a}\|_2, \varepsilon)$ denotes numerically stable ℓ_2 normalization. Missing views are not passed through the encoders or decoders; their representations are instead recovered by the subsequent retrieval module in the shared latent space.

3.2 Target-Conditioned Direct-Context Retrieval

For a mini-batch $\mathcal{B}$, we construct a batch-local evidence bank from all observed sample–view entries:

$$\mathcal{I}_{\mathcal{B}} = \{(j,u) \mid j \in \mathcal{B},\, m_j^{(u)} = 1\}, \qquad \mathbf{k}_j^{(u)} = \nu\left(g_k([\mathbf{z}_j^{(u)}; \mathbf{e}_u])\right). \tag{2}$$

Here, $\mathbf{e}_u \in \mathbb{R}^d$ is a learnable view embedding, $g_k : \mathbb{R}^{2d} \to \mathbb{R}^d$ is the shared key network, and $[\cdot;\cdot]$ denotes concatenation. Each $(j,u) \in \mathcal{I}_{\mathcal{B}}$ indexes a key–value entry $(\mathbf{k}_j^{(u)}, \mathbf{z}_j^{(u)})$. The bank is reconstructed for each mini-batch and is not retained across batches.

For a missing target view $v \notin \mathcal{O}_i$, its query is constructed from the available-view summary and the target-view embedding:

$$\mathbf{a}_i = \nu\left(\frac{1}{|\mathcal{O}_i|}\sum_{u\in\mathcal{O}_i}\mathbf{z}_i^{(u)}\right), \qquad \mathbf{q}_i^{(v)} = \nu(g_q([\mathbf{a}_i;\mathbf{e}_v])), \tag{3}$$

where $g_q : \mathbb{R}^{2d} \to \mathbb{R}^d$ is the shared query network.

For each candidate $(j,u) \in \mathcal{I}_{\mathcal{B}}$, we compute

$$s_{ij}^{(v,u)} = \frac{\langle \mathbf{q}_i^{(v)}, \mathbf{k}_j^{(u)}\rangle}{\tau_r} + \beta_s \mathbf{1}[i=j] + \beta_r R_{v,u}, \tag{4}$$

where τ_r is the retrieval temperature, β_s and β_r control the same-instance and cross-view relation biases, respectively, and $R \in \mathbb{R}^{V\times V}$ is a learnable view-relation matrix. For a missing target view, the same-instance term can only favor its other observed views.

Select the $k = \min(K_r, |\mathcal{I}_{\mathcal{B}}|)$ highest-scoring entries, indexed by (j_t, u_t), and write $s_{it}^{(v)} := s_{ij_t}^{(v,u_t)}$. Then,

$$w_{it}^{(v)} = \frac{\exp s_{it}^{(v)}}{\sum_{\ell=1}^{k}\exp s_{i\ell}^{(v)}}, \qquad \widetilde{\mathbf{z}}_i^{(v)} = \nu\left(\sum_{t=1}^{k} w_{it}^{(v)}\mathbf{z}_{j_t}^{(u_t)}\right). \tag{5}$$

Thus, the target-conditioned query determines evidence allocation, while the missing representation is recovered directly from observed latent values. For a fixed retrieved set, the softmax weights admit an entropy-regularized allocation interpretation, and the pre-normalized context lies in the convex hull of the selected evidence; formal statements are provided in Appendix C.

We estimate retrieval reliability from weight concentration and evidence agreement:

$$\begin{gathered} c_{i,\mathrm{ent}}^{(v)} = 1 + \frac{\sum_{t=1}^{k} w_{it}^{(v)}\log w_{it}^{(v)}}{\log k}, \quad k>1, \qquad c_{i,\mathrm{agr}}^{(v)} = \frac{1+\sum_{t=1}^{k} w_{it}^{(v)}\rho_{it}^{(v)}}{2} \\ c_i^{(v)} = g_c\Big([c_{i,\mathrm{ent}}^{(v)}; c_{i,\mathrm{agr}}^{(v)}]\Big) \end{gathered} \tag{6}$$

where $\rho_{it}^{(v)} = \langle \mathbf{z}_{j_t}^{(u_t)}, \widetilde{\mathbf{z}}_i^{(v)}\rangle$ is the agreement between the t-th retrieved entry and the recovered representation. The concentration score $c_{i,\mathrm{ent}}^{(v)}$ increases as the retrieval weights become more concentrated and is set to 1 when $k=1$, whereas $c_{i,\mathrm{agr}}^{(v)}$ increases when the retrieved evidence is more coherent with the recovered representation. After pretraining, the sigmoid network g_c combines the two statistics into the final reliability weight. Observed views are assigned $c_i^{(v)} = 1$, and the resulting weights are used only in the cross-view consensus objective.

3.3 Observed-View Enhancement via Self-Excluding Retrieval

Direct replacement is suitable for a missing view but may overwrite reliable information in an observed representation. Therefore, for an observed target view $v \in \mathcal{O}_i$ with $|\mathcal{O}_i| > 1$, we construct a query using only the other observed views:

$$\mathbf{a}_{i,-v} = \nu\left(\frac{1}{|\mathcal{O}_i|-1}\sum_{u\in\mathcal{O}_i\setminus\{v\}}\mathbf{z}_i^{(u)}\right), \qquad \mathbf{q}_{i,-v}^{(v)} = \nu(g_q([\mathbf{a}_{i,-v};\mathbf{e}_v])). \tag{7}$$

Unlike missing-view recovery, the candidate set contains only the target-view representations of other observed instances:

$$\mathcal{I}_{\mathcal{B}}^{(v,i)} = \Big\{ j \in \mathcal{B}\setminus\{i\} \mid m_j^{(v)} = 1\Big\}, \qquad \bar{s}_{ij}^{(v)} = \frac{\langle \mathbf{q}_{i,-v}^{(v)}, \mathbf{k}_j^{(v)}\rangle}{\tau_r}. \tag{8}$$

Let $\mathcal{T}_{i,\mathrm{obs}}^{(v)}$ contain the $\min(K_r, |\mathcal{I}_{\mathcal{B}}^{(v,i)}|)$ highest-scoring candidates, and let $\bar{w}_{ij}^{(v)}$ denote their softmax weights. The observed representation is refined through a residual update:

$$\widetilde{\mathbf{z}}_i^{(v)} = \nu\left(\mathbf{z}_i^{(v)} + \alpha\,\nu\left(\sum_{j\in\mathcal{T}_{i,\mathrm{obs}}^{(v)}} \bar{w}_{ij}^{(v)}\mathbf{z}_j^{(v)}\right)\right), \tag{9}$$

where α controls the contribution of the retrieved context. If the candidate set is empty or no other view of instance i is observed, we retain $\widetilde{\mathbf{z}}_i^{(v)} = \mathbf{z}_i^{(v)}$. Both queries and values use the original encoded observations; completed and refined outputs are never reused as retrieval values. The missing-view and observed-view paths share the view embeddings and query/key networks but employ different candidate sets and update rules.

Multi-view fusion. After recovering missing views and refining eligible observed views, we obtain the final clustering representation by

$$\mathbf{f}_i = \nu\left(\frac{1}{V}\sum_{v=1}^{V}\widetilde{\mathbf{z}}_i^{(v)}\right). \tag{10}$$

3.4 Representation Learning and Balanced Consensus

We optimize the model using three objectives. First, reconstruction preserves the information contained in the observed inputs:

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{V}\sum_{v=1}^{V}\frac{\sum_{i\in\mathcal{B}} m_i^{(v)}\left\|\mathbf{x}_i^{(v)} - \widehat{\mathbf{x}}_i^{(v)}\right\|_2^2}{\max\left(1, \sum_{i\in\mathcal{B}} m_i^{(v)}\right)}. \tag{11}$$

This objective is evaluated only on observed views using the reconstructions defined in Eq. equation 1.

Second, a symmetric InfoNCE objective (Oord et al., 2018) aligns each completed view with the fused representation. Let $B = |\mathcal{B}|$, $F = [\mathbf{f}_i^\top]_{i\in\mathcal{B}}$, and $\widetilde{Z}^{(v)} = [\widetilde{\mathbf{z}}_i^{(v)\top}]_{i\in\mathcal{B}}$. The global-view alignment objective is

$$\mathcal{L}_{\mathrm{GVA}} = \frac{1}{2V}\sum_{v=1}^{V}\left[\mathrm{CE}\left(\frac{F\widetilde{Z}^{(v)\top}}{\tau_a}, \mathbf{t}\right) + \mathrm{CE}\left(\frac{\widetilde{Z}^{(v)}F^\top}{\tau_a}, \mathbf{t}\right)\right], \tag{12}$$

where $\mathbf{t} = (1, \ldots, B)$ contains the matching indices, i.e., $t_i = i$. Thus, the two terms perform fused-to-view and view-to-fused alignment, respectively, using the same-instance pair as the positive match.

Third, after representation pretraining, K-means centers initialize C trainable prototypes $\{\mathbf{p}_c\}_{c=1}^{C}$. Each completed view produces prototype logits $\ell_{ic}^{(v)} = \langle\widetilde{\mathbf{z}}_i^{(v)}, \nu(\mathbf{p}_c)\rangle$. Writing $L^{(v)} = [\ell_{ic}^{(v)}]$, we obtain the balanced assignments as $Q^{(v)} = \mathrm{SK}(\mathrm{stopgrad}(L^{(v)}); \epsilon)$, where SK denotes Sinkhorn normalization. Following balanced online assignments (Caron et al., 2020), each row of $Q^{(v)}$ sums to one, while its batch-level cluster marginal is approximately uniform.

For each ordered view pair $v \neq u$, the balanced assignments of view v supervise the predictions of view u:

$$\begin{aligned}\mathcal{L}_{v\to u} &= \frac{\sum_{i\in\mathcal{B}} c_i^{(v)}c_i^{(u)}\left[-\sum_{c=1}^{C} Q_{ic}^{(v)}\log\left(\mathrm{softmax}(\boldsymbol{\ell}_i^{(u)}/\tau_c)\right)_c\right]}{\sum_{i\in\mathcal{B}} c_i^{(v)}c_i^{(u)}},\\ \mathcal{L}_{\mathrm{CVC}} &= \frac{1}{V(V-1)}\sum_{v\neq u}\mathcal{L}_{v\to u}.\end{aligned} \tag{13}$$

The reliability product $c_i^{(v)}c_i^{(u)}$ reduces the contribution of uncertain completed views. Gradients do not pass through the Sinkhorn targets, while the prediction branch, prototypes, and reliability network remain trainable.

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_a \mathcal{L}_{\text{GVA}} + \lambda_c \mathcal{L}_{\text{CVC}}, \tag{14}$$

where λ_a and λ_c control the contributions of global-view alignment and cross-view consensus, respectively. During representation pretraining, λ_c is set to zero; the prototypes are then initialized and the cross-view consensus objective is enabled.

Optimization and inference. Training proceeds in two stages. During the first T_p epochs, retrieval-based recovery and observed-view enhancement are active, while $\lambda_c = 0$; the model is optimized using $\mathcal{L}_{\text{rec}} + \lambda_a \mathcal{L}_{\text{GVA}}$. After pretraining, we apply K-means to the fused representations of the full dataset and use the resulting centers to initialize the trainable prototypes. The reliability network and cross-view consensus objective are then activated, and the model is optimized using the full objective in Eq. equation 14. At inference, the final fused representations are clustered using K-means, while prototype assignments are used only during training.

Computational complexity. Let $B = |\mathcal{B}|$ and $M_{\mathcal{B}} = |\mathcal{I}_{\mathcal{B}}| \leq BV$ denote the batch size and the number of observed entries in the batch-local evidence bank, respectively. For Q_m missing-view queries, the dominant similarity computation of missing-view recovery requires $O(Q_m M_{\mathcal{B}} d)$ time. Observed-view enhancement requires $O(\sum_{v=1}^{V} Q_v B_v d)$ time, where Q_v and B_v denote the numbers of eligible queries and candidate entries for view v, respectively. Since $Q_v, B_v \leq B$, this term is upper-bounded by $O(VB^2 d)$. Therefore, the overall retrieval complexity per mini-batch is

$$O\left(Q_m M_{\mathcal{B}} d + \sum_{v=1}^{V} Q_v B_v d\right), \tag{15}$$

and storing the corresponding keys and values requires $O(M_{\mathcal{B}} d)$ memory.

4 Experiments

We evaluate TCRIMVC through baseline comparisons, experiments across missing rates, ablation and hyperparameter studies, representation visualizations, and qualitative analysis of missing-view recovery.

4.1 Experimental Setup

We evaluate our method on four widely used multi-view datasets. MNIST-10k comprises 10,000 handwritten digit images from 10 classes, with ISO, LDA, and NPE features forming three views. Reuters-1200 contains 1,200 documents from 6 classes, represented by five language views: English, French, German, Spanish, and Italian. NGs includes 500 news documents from 5 classes with three feature views, while Scene-15 consists of 4,485 images from 15 scene categories represented by GIST, PHOG, and LBP features. We assess clustering performance using clustering accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (ARI), with higher values indicating better performance.

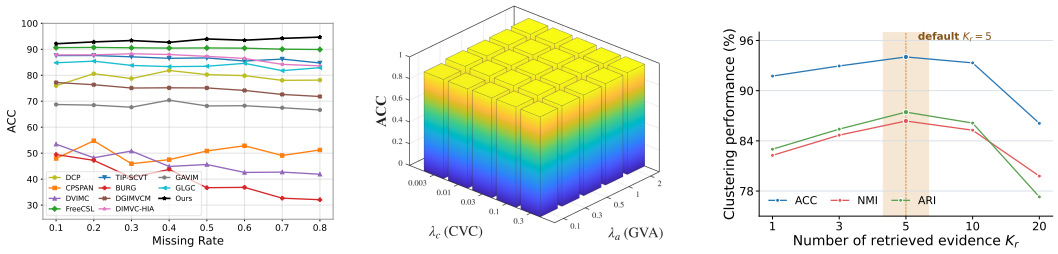
All experiments are implemented in PyTorch and conducted on an NVIDIA GeForce RTX 3090 GPU. We use Adam with a learning rate of 3×10^{-3} and a batch size of 256. Based on parameter sensitivity analysis, we set $K_r = 5$, $\tau_r = 0.2$, $\alpha = 0.2$, and $\lambda_c = 0.1$, and activate the cross-view clustering-consistency objective after pretraining. Final cluster assignments are obtained by applying K-means to the fused representations. We report the mean results over five independent runs at the final training epoch, with the corresponding standard deviations provided in the supplementary material due to space constraints.

4.2 Comparison with Existing Methods

We compare TCRIMVC with ten representative IMVC methods: DGIMVCM (Zhang et al., 2026), GLGC (He et al., 2026), GAVIM (Chen et al., 2026), DIMVC-HIA (Du et al., 2026), DCP (Lin et al., 2022), CPSPAN (Jin et al., 2023), DVIMC (Xu et al., 2024), FreeCSL (Dai et al., 2025), TIP-SCVT (Dong et al., 2025), and BURG (Jin et al., 2025). These methods cover several representative

Table 1: Clustering performance (%) on four benchmark datasets (incomplete vs. complete). The best and second-best results per column are bolded and underlined, respectively.

Setting	Method	MNIST-10k			Reuters-1200			NGs			Scene-15		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Incomplete	DGIMVCM	73.75	68.57	60.42	31.43	13.34	6.10	30.08	4.56	2.87	<u>44.58</u>	<u>42.80</u>	<u>27.24</u>
	GLGC	83.95	70.56	68.85	35.70	12.95	11.05	53.76	28.72	23.85	35.11	33.71	19.40
	GAVIM	68.28	64.79	56.90	47.38	28.96	18.34	27.92	4.42	2.00	39.26	34.13	21.03
	DIMVC-HIA	87.00	75.26	74.09	44.33	22.46	17.33	<u>60.64</u>	<u>34.66</u>	<u>29.91</u>	34.32	35.27	18.55
	DCP	80.25	72.07	68.59	26.58	11.78	5.31	21.44	3.54	4.00	38.93	40.88	22.77
	CPSPAN	50.97	55.91	40.15	31.67	13.45	9.16	31.20	13.90	7.34	40.98	41.21	24.02
	DVIMC	45.64	32.03	23.19	37.17	13.71	9.14	28.28	7.71	4.97	26.51	19.14	9.06
	FreeCSL	<u>90.43</u>	<u>79.84</u>	<u>80.63</u>	44.42	21.23	16.45	36.80	11.34	6.85	34.83	33.87	19.13
	TIP-SCVT	86.60	74.89	73.71	<u>52.50</u>	<u>31.07</u>	<u>22.75</u>	46.60	25.91	14.72	31.75	31.37	16.41
	BURG	52.23	46.62	25.76	37.17	20.44	13.64	20.60	1.55	0.00	27.56	27.74	11.03
	Ours	**94.03**	**86.37**	**87.44**	**65.52**	**42.55**	**36.97**	**89.36**	**73.82**	**75.45**	**55.01**	**52.07**	**39.10**
Complete	DGIMVCM	78.62	76.32	70.17	31.02	13.65	7.03	33.72	10.70	7.13	<u>50.45</u>	**50.71**	<u>33.63</u>
	GLGC	85.05	72.03	70.65	33.30	13.08	11.18	<u>76.56</u>	<u>57.71</u>	<u>55.32</u>	34.99	35.48	19.82
	GAVIM	69.34	67.63	59.22	45.60	28.45	17.88	36.60	14.27	9.43	42.42	39.04	24.84
	DIMVC-HIA	86.02	74.10	72.44	45.92	24.70	18.99	55.76	41.64	29.95	26.78	26.98	12.02
	DCP	78.46	69.41	64.26	34.67	19.49	9.96	21.52	3.15	2.00	40.02	44.47	24.35
	CPSPAN	48.90	54.83	38.66	32.42	13.02	9.24	30.60	15.04	4.58	40.91	40.11	23.35
	DVIMC	54.24	39.19	29.85	36.25	18.30	11.20	34.20	16.36	12.06	32.11	29.18	14.95
	FreeCSL	<u>90.86</u>	<u>80.52</u>	<u>81.31</u>	49.33	<u>29.46</u>	<u>23.45</u>	45.20	21.99	15.92	39.60	41.27	23.24
	TIP-SCVT	87.34	75.50	74.93	<u>51.58</u>	28.67	20.99	34.00	30.25	8.18	36.81	35.81	20.62
	BURG	75.67	74.58	68.77	46.83	28.97	19.88	32.40	8.29	4.15	28.34	34.37	15.03
	Ours	**91.94**	**82.57**	**83.26**	**64.60**	**40.72**	**35.51**	**97.32**	**92.03**	**93.36**	**52.71**	<u>49.21</u>	**35.77**



(a) ACC vs. missing rate

(b) Parameter sensitivity-λ_c and λ_a

(c) Parameter sensitivity- K_r

Figure 2: Robustness and parameter sensitivity results on MNIST-10k.

paradigms, including graph-based modeling, missing-view completion, generative recovery, and contrastive representation learning. Table 1 reports the results under the incomplete setting ($\eta = 0.5$) and the complete setting ($\eta = 0$). Across the 24 dataset–metric comparisons, TCRIMVC achieves the best result in 23 cases and ranks second in the remaining case. As shown in Table 1, under the incomplete setting, TCRIMVC achieves the best ACC, NMI, and ARI across all four datasets. Compared with the strongest baseline on each dataset, TCRIMVC improves NMI by 6.53, 11.48, 39.16, and 9.27 percentage points on MNIST-10k, Reuters-1200, NGs, and Scene-15, respectively, alongside consistent gains in ACC and ARI. The particularly large improvements on Reuters-1200 and NGs highlight its effectiveness under incomplete observations. Under the complete setting, TCRIMVC achieves the best results with substantial gains on Reuters-1200 and NGs. Overall, these results demonstrate the consistent effectiveness of TCRIMVC across datasets with diverse scales, view configurations, and modalities.

To evaluate the robustness of TCRIMVC under different levels of view incompleteness, we compare its ACC with recent state-of-the-art methods on MNIST-10k as the missing rate η varies from 0.1 to 0.8 in increments of 0.1. As shown in Fig. 2(a), TCRIMVC achieves the highest ACC at every tested missing rate and maintains consistently strong performance as η increases. At $\eta = 0.8$, TCRIMVC outperforms FreeCSL, the strongest baseline at this setting, by 4.77 percentage points. Together with the main results at $\eta = 0$ and $\eta = 0.5$, these results demonstrate the robustness of TCRIMVC across different levels of view incompleteness.

Table 2: Ablation results (%) on MNIST-10k at $\eta = 0.5$.

Components		Loss Terms			Metrics		
Recovery	Enhancement	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{GVA}}$	$\mathcal{L}_{\text{CVC}}$	**ACC**	**NMI**	**ARI**
–	–	✓	✓	✓	85.94	75.74	74.44
✓	–	✓	✓	✓	91.51	81.71	82.47
✓	✓	–	✓	✓	90.81	81.64	81.91
✓	✓	✓	–	✓	82.67	76.27	71.97
✓	✓	✓	✓	–	91.55	85.46	85.33
✓	✓	✓	✓	✓	**94.03**	**86.37**	**87.44**

Table 3: Runtime (s) and ACC (%) comparison results.

		DGIMVCM		GLGC		GAVIM		DIMVC-HIA		TCRIMVC (Ours)	
Missing rate	Dataset	Time (s)	ACC (%)	Time (s)	ACC (%)	Time (s)	ACC (%)	Time (s)	ACC (%)	Time (s)	ACC (%)
0.3	MNIST-10k	547.31	75.09±1.24	588.11	83.79±1.71	1637.60	67.69±1.10	5500.41	88.28±0.31	**504.72**	**93.36±0.77**
	NGs	31.77	31.73±1.40	22.81	71.67±9.05	97.72	43.73±11.38	241.98	64.33±7.56	**17.34**	**93.87±0.23**
0.7	MNIST-10k	541.99	72.61±1.92	490.18	81.77±6.82	1658.69	67.48±1.01	6087.20	84.19±4.65	**483.39**	**94.21±0.44**
	NGs	29.77	27.60±1.44	23.77	52.13±9.06	100.30	31.40±4.90	280.81	44.00±4.13	**16.67**	**87.93±0.23**

4.3 Model Analysis

Parameter Sensitivity. Figures 2(b) and 2(c) examine the effects of the loss weights (λ_a, λ_c) and the retrieval size K_r, respectively. The adopted setting $(\lambda_a, \lambda_c) = (1, 0.1)$ lies in a high-performing region, while neighboring combinations produce comparable ACC values, indicating limited sensitivity to moderate changes in the loss weights. ACC, NMI, and ARI increase as K_r grows from 1 to 5 and reach their highest values at $K_r = 5$, whereas larger values degrade performance by introducing less relevant evidence. Accordingly, we use $K_r = 5$ in all clustering experiments.

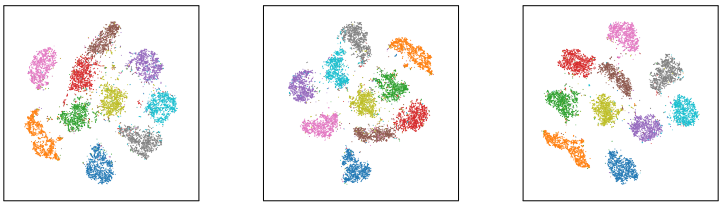
Ablation Study. Table 2 evaluates the contributions of retrieval-based recovery, observed-view enhancement, and the three training objectives on MNIST-10k at $\eta = 0.5$. When observed-view enhancement is disabled, retrieval-based recovery improves ACC, NMI, and ARI over the cross-view MLP predictor by 5.57, 5.97, and 8.03 percentage points, respectively. Further enabling observed-view enhancement yields additional gains of 2.52, 4.66, and 4.97 points. Among the loss terms, removing GVA causes the largest reductions of 11.36, 10.10, and 15.47 points, while removing reconstruction or CVC also consistently degrades performance. These results demonstrate that the retrieval modules and training objectives make complementary contributions to the full model.

Visualization Analysis. Figure 3 compares the learned fused representations of three model variants. Without retrieval, samples from the same class are relatively dispersed and different classes exhibit substantial overlap. Retrieval-based recovery produces more compact class structures, although some neighboring classes remain partially mixed. The full model forms the most coherent clusters and the clearest inter-class boundaries, supporting the complementary contributions of retrieval-based recovery and observed-view enhancement observed in the ablation results.

Computational Efficiency. To complement the theoretical complexity analysis, we evaluate the practical efficiency of TCRIMVC by comparing its wall-clock training time and ACC with four representative baselines on MNIST-10k and NGs at $\eta = 0.3$ and $\eta = 0.7$. As shown in Table 3, TCRIMVC consistently achieves the highest ACC and the shortest runtime across all four settings. Compared with the strongest baseline in each setting, it improves ACC by 5.08–35.80 percentage points while achieving a $1.01\times$–$1.43\times$ speedup over the fastest competing method. These results confirm the practical efficiency of TCRIMVC and show that its performance gains do not come at the cost of excessive computational overhead.

4.4 Cross-view Image Completion

To further examine whether the proposed retrieval mechanism captures transferable cross-view correspondence, we conduct an auxiliary image completion experiment on the paired RGB and sketch



(a) w/o Retrieval (b) w/o Obs. Enh. (c) Full

Figure 3: t-SNE visualizations of the fused representations on MNIST-10k at $\eta = 0.5$.

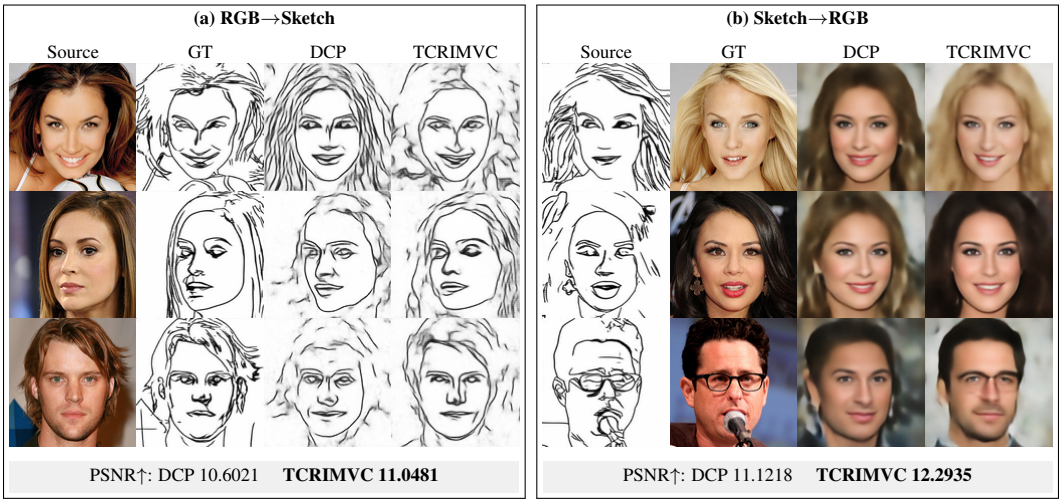


Figure 4: Cross-view image completion on MM-CelebA-HQ at $\eta = 0.3$. The columns show the source view, ground-truth (GT) target, and the results of DCP and TCRIMVC, with dataset-level PSNR reported below.

views of MM-CelebA-HQ (Xia et al., 2021). We evaluate both RGB→Sketch and Sketch→RGB completion at $\eta = 0.3$ and compare TCRIMVC with DCP (Lin et al., 2022) using identical missing masks and evaluation splits. The retrieved representations are decoded by the corresponding target-view decoder, and completion quality is evaluated using PSNR, where a higher value indicates better performance.

As shown in Fig. 4, TCRIMVC outperforms DCP in both completion directions, improving PSNR by 0.45 dB for RGB→Sketch and 1.17 dB for Sketch→RGB. The qualitative results also show that TCRIMVC better preserves the facial structures and semantic attributes of the target views, producing outputs that more closely match the ground truth. These results indicate that the proposed retrieval mechanism can capture meaningful cross-view correspondence beyond the clustering task.

5 Conclusion

In this paper, we presented a retrieval-based framework (TCRIMVC) that reformulates incomplete multi-view representation learning as target-conditioned direct-context retrieval. By directly retrieving observed latent evidence, TCRIMVC unifies missing-view recovery and observed-view enhancement without requiring an additional representation generator. Reliability-aware cross-view consensus further mitigates the impact of uncertain completions on clustering. Extensive experiments across multiple datasets and missing rates demonstrate the effectiveness and robustness of TCRIMVC, while ablation, visualization, and efficiency analyses validate its key designs. Moreover, cross-view image completion experiments demonstrate the transferability of the learned representations beyond clustering.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing and to provide feedback on methodological exposition and interpretation of results. The authors take responsibility for the final content of the paper.

REFERENCES

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Guoqing Chao, Yi Jiang, and Dianhui Chu. Incomplete contrastive multi-view clustering with high-confidence guiding. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 11221–11229, 2024.

Wenlan Chen, Lu Gao, Daoyuan Wang, Fei Guo, and Cheng Liang. Geometry-aware variational information maximization for deep incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 20289–20297, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5071–5081. IEEE, 2025.

Xiaojian Ding, Lin Zhao, Xian Li, and Xiaoying Zhu. Incomplete multi-view clustering via hierarchical semantic alignment and cooperative completion. *Advances in Neural Information Processing Systems*, 38:52175–52195, 2026.

Zhibin Dong, Dayu Hu, Jiaqi Jin, Siwei Wang, Xinwang Liu, and En Zhu. Selective cross-view topology for deep incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 34:4792–4805, 2025.

Yiming Du, Ziyu Wang, Jian Li, Rui Ning, and Lusi Li. Deep incomplete multi-view clustering via hierarchical imputation and alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 20941–20949, 2026.

Xin Gao and Jian Pu. Deep incomplete multi-view learning via cyclic permutation of vaes. In *International Conference on Learning Representations*, volume 2025, pp. 99609–99643, 2025.

Hongqing He, Jie Xu, Wenyuan Yang, Yonghua Zhu, Guoqiu Wen, and Xiaofeng Zhu. Global-graph guided and local-graph weighted contrastive learning for unified clustering on incomplete and noise multi-view data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24813–24822, 2026.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11600–11609. IEEE, 2023.

Jiaqi Jin, Siwei Wang, Zhibin Dong, Xihong Yang, Xinwang Liu, En Zhu, and Kunlun He. Deep incomplete multi-view clustering with distribution dual-consistency recovery guidance. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1016–1026. IEEE, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. Incomplete multi-view clustering via prototype-based imputation. *arXiv preprint arXiv:2301.11045*, 2023.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11169–11178. IEEE, 2021.

Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2022.

Suyuan Liu, Siwei Wang, Ke Liang, Junpu Zhang, Zhibin Dong, Tianrui Liu, En Zhu, Kunlun He, and Xinwang Liu. Alleviate anchor-shift: Explore blind spots with cross-view reconstruction for incomplete multi-view clustering. *Advances in Neural Information Processing Systems*, 37: 87509–87531, 2024.

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

Emanuele Palumbo, Laura Manduchi, Sonia Laguna, Daphné Chopard, and Julia E Vogt. Deep generative clustering with multimodal diffusion variational autoencoders. In *International Conference on Learning Representations*, volume 2024, pp. 24286–24310, 2024.

Huayi Tang and Yong Liu. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International conference on machine learning*, pp. 21090–21110. PMLR, 2022.

Qianqian Wang. Llm-damvc: A large language model assisted dynamic agent for multi-view clustering. *Advances in Neural Information Processing Systems*, 38:119515–119534, 2026.

Qianqian Wang, Haiming Xu, Wei Feng, and Quanxue Gao. Adversarial fair incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10011–10019, 2026a.

Yiming Wang, Qun Li, Dongxia Chang, Jie Wen, Hua Dai, Fu Xiao, and Yao Zhao. Learning from disjoint views: A contrastive prototype matching network for fully incomplete multi-view clustering. *Advances in Neural Information Processing Systems*, 38:18265–18284, 2026b.

Ziyu Wang, Yiming Du, Yao Wang, Rui Ning, and Lusi Li. Deep incomplete multi-view clustering via multi-level imputation and contrastive alignment. *Neural Networks*, 181:106851, 2025.

Benyu Wu, Wei Du, Jun Wang, and Guoxian Yu. Imputation-free incomplete multi-view clustering via knowledge distillation. In *IJCAI*, pp. 6570–6578, 2025.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Towards open-world text-guided face image generation and manipulation. *arXiv preprint arXiv:2104.08910*, 2021.

Gehui Xu, Jie Wen, Chengliang Liu, Bing Hu, Yicheng Liu, Lunke Fei, and Wei Wang. Deep variational incomplete multi-view clustering: Exploring shared clustering structures. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 16147–16155, 2024.

Jie Xu, Chao Li, Yazhou Ren, Liang Peng, Yujie Mo, Xiaoshuang Shi, and Xiaofeng Zhu. Deep incomplete multi-view clustering via mining cluster complementarity. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 8761–8769, 2022.

Jie Xu, Chao Li, Liang Peng, Yazhou Ren, Xiaoshuang Shi, Heng Tao Shen, and Xiaofeng Zhu. Adaptive feature projection with distribution alignment for deep incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 32:1354–1366, 2023.

Weiqing Yan, Meiqi Gu, Jinlai Ren, Guanghui Yue, Zhaowei Liu, Jindong Xu, and Weisi Lin. Collaborative structure and feature learning for multi-view clustering. *Information Fusion*, 98: 101832, 2023a.

Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19863–19872. IEEE, 2023b.

Weiqing Yan, Yuanyang Zhang, Chang Tang, Wujie Zhou, and Weisi Lin. Anchor-sharing and cluster-wise contrastive network for multiview representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3797–3807, 2024.

Yuanyang Zhang, Yijie Lin, Weiqing Yan, Li Yao, Xinhang Wan, Guangyuan Li, Chao Zhang, Guanzhou Ke, and Jie Xu. Incomplete multi-view clustering via diffusion contrastive generation. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 22650–22658, 2025.

Zhenghao Zhang, Jun Xie, Xingchen Chen, Tao Yu, Hongzhu Yi, Kaixin Xu, Yuanxiang Wang, Tianyu Zong, Xinming Wang, Jiahuan Chen, et al. Dynamic deep graph learning for incomplete multi-view clustering with masked graph reconstruction loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 28600–28608, 2026.

Xinying Zhao, Weiqing Yan, Jinlai Ren, Jindong Xu, Zhaowei Liu, Guanghui Yue, and Chang Tang. Graph-filtering and high-order bipartite graph based multiview graph clustering. *Digital Signal Processing*, 133:103847, 2023.

Taichun Zhou, Zhibin Dong, Hao Tan, Siwei Wang, Xinwang Liu, En Zhu, Di Hu, Tianrui Liu, Chuankun Li, and Kunlun He. Imbalanced view contribution evaluation and refinement for deep incomplete multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39606–39616, 2026.

Deyin Zhuang, Jian Dai, Xingfeng Li, Xi Wu, Yuan Sun, and Zhenwen Ren. Robust graph contrastive learning for incomplete multi-view clustering. In *IJCAI*, pp. 7282–7290, 2025.

A Additional Experimental Results

A.1 Complete Results with Standard Deviations

To provide a statistically consistent comparison with recent methods and evaluate TCRIMVC beyond the main benchmarks, Tables 4 and 5 report the mean and standard deviation over five independent runs at $\eta = 0$ and $\eta = 0.5$. Table 4 supplements the main comparison with complete five-run results for four recent 2026 methods on the four primary datasets, while Table 5 extends the same evaluation protocol to LandUse21 and NUS-WIDE.

Table 4: Five-run comparison with recent methods on the four main datasets. Results are reported as mean±standard deviation (%). The best and second-best results are shown in bold and underlined, respectively.

		$\eta = 0$			$\eta = 0.5$		
Dataset	Method	ACC	NMI	ARI	ACC	NMI	ARI
MNIST-10k	DGIMVCM	78.62±0.79	76.32±1.28	70.17±1.69	73.75±2.03	68.57±1.58	60.42±2.86
	GLGC	85.05±0.92	72.03±0.99	70.65±1.56	83.95±2.47	70.56±2.31	68.85±3.59
	GAVIM	69.34±0.96	67.63±0.66	59.22±0.62	68.28±1.08	64.79±0.77	56.90±0.73
	DIMVC-HIA	86.02±1.27	74.10±1.62	72.44±2.27	87.00±0.82	75.26±0.90	74.09±1.47
	TCRIMVC	**91.94±0.20**	**82.57±0.29**	**83.26±0.40**	**94.03±0.46**	**86.37±0.70**	**87.44±0.94**
Reuters-1200	DGIMVCM	31.02±4.86	13.65±3.07	7.03±2.67	31.43±2.43	13.34±1.70	6.10±1.66
	GLGC	33.30±1.50	13.08±2.35	11.18±2.01	35.70±5.42	12.95±2.31	11.05±2.34
	GAVIM	45.60±3.85	28.45±1.80	17.88±1.80	47.38±2.07	28.96±1.79	18.34±1.81
	DIMVC-HIA	45.92±3.11	24.70±3.17	18.99±3.32	44.33±3.26	22.46±2.35	17.33±2.36
	TCRIMVC	**64.60±0.87**	**40.72±0.92**	**35.51±1.02**	**65.52±1.47**	**42.55±1.97**	**36.97±1.58**
NGs	DGIMVCM	33.72±0.80	10.70±0.51	7.13±0.52	30.08±0.88	4.56±0.77	2.87±0.67
	GLGC	76.56±8.14	57.71±10.57	55.32±13.28	53.76±4.68	28.72±4.20	23.85±4.74
	GAVIM	36.60±2.59	14.27±3.34	9.43±2.73	27.92±1.74	4.42±1.27	2.00±0.93
	DIMVC-HIA	55.76±9.18	41.64±9.22	29.95±9.74	60.64±6.94	34.66±7.47	29.91±7.56
	TCRIMVC	**97.32±1.43**	**92.03±3.29**	**93.36±3.51**	**89.36±1.66**	**73.82±3.03**	**75.45±3.44**
Scene-15	DGIMVCM	50.45±1.52	**50.71±0.82**	33.63±1.59	44.58±1.16	42.80±0.47	27.24±0.41
	GLGC	34.99±0.91	35.48±0.28	19.82±0.47	35.11±0.60	33.71±0.75	19.40±0.21
	GAVIM	42.42±1.77	39.04±1.13	24.84±1.47	39.26±2.79	34.13±1.12	21.03±1.33
	DIMVC-HIA	26.78±1.80	26.98±2.59	12.02±1.37	34.32±1.17	35.27±1.46	18.55±1.18
	TCRIMVC	**52.71±1.20**	49.21±1.42	**35.77±1.43**	**55.01±1.93**	**52.07±2.12**	**39.10±2.38**

Table 5: Five-run comparison on two additional datasets. Results are reported as mean±standard deviation (%). The best and second-best results are shown in bold and underlined, respectively.

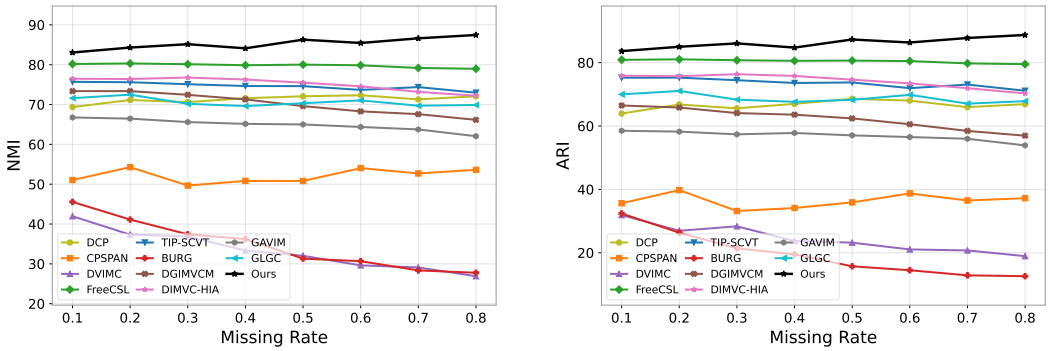
		$\eta = 0$			$\eta = 0.5$		
Dataset	Method	ACC	NMI	ARI	ACC	NMI	ARI
LandUse21	DGIMVCM	30.76±0.50	**38.57±0.63**	16.46±0.67	26.48±0.85	30.36±1.00	12.07±0.72
	GLGC	23.65±1.36	27.24±1.77	9.80±1.12	23.53±1.35	26.40±1.82	9.83±1.26
	GAVIM	21.91±1.08	25.27±1.10	8.33±0.90	22.13±1.85	25.15±1.17	8.68±1.00
	DIMVC-HIA	16.60±1.28	19.29±1.18	4.71±0.73	21.42±2.07	22.96±0.96	7.76±1.11
	TCRIMVC	**32.15±1.32**	37.07±0.86	**17.94±0.80**	**30.02±0.58**	**33.21±1.05**	**15.20±0.73**
NUS-WIDE	DGIMVCM	28.74±1.10	14.98±0.52	9.39±0.32	25.10±1.64	12.30±0.70	6.99±0.91
	GLGC	26.03±1.43	13.90±0.45	8.38±0.80	25.97±1.90	12.99±1.02	8.06±1.00
	GAVIM	23.77±1.17	13.26±0.57	6.49±0.35	22.90±1.15	11.58±0.75	5.76±0.56
	DIMVC-HIA	22.49±1.58	12.14±0.93	5.37±0.80	24.35±1.09	12.52±0.46	6.55±0.35
	TCRIMVC	**30.87±0.94**	**17.29±0.37**	**11.30±0.70**	**30.53±1.30**	**16.94±0.75**	**11.41±1.00**

Across the two tables, TCRIMVC achieves the best result in 34 of the 36 dataset–setting–metric comparisons and ranks second in the remaining two cases, corresponding to NMI on Scene-15 and LandUse21 at $\eta = 0$. In particular, it ranks first in all 18 comparisons under the incomplete setting. At $\eta = 0.5$, TCRIMVC improves ACC, NMI, and ARI over the strongest baselines by 3.54, 2.85, and 3.13 percentage points on LandUse21, and by 4.56, 3.95, and 3.35 points on NUS-WIDE,

respectively. These results, together with the relatively small standard deviations, demonstrate that the performance advantage of TCRIMVC remains stable across independent runs and generalizes to datasets beyond the main comparison.

A.2 Additional Results under Different Missing Rates

The main paper reports the MNIST-10k ACC curve. To complete the evaluation, Fig. 5 reports its NMI and ARI curves, while Figs. 6–8 report all three metrics on Reuters-1200, NGs, and Scene-15. Every curve uses the same requested missing settings $\eta \in \{0.1, 0.2, \ldots, 0.8\}$ and three independent runs.



(a) NMI (b) ARI

Figure 5: NMI and ARI on MNIST-10k under different requested missing settings. ACC is reported in the main paper.

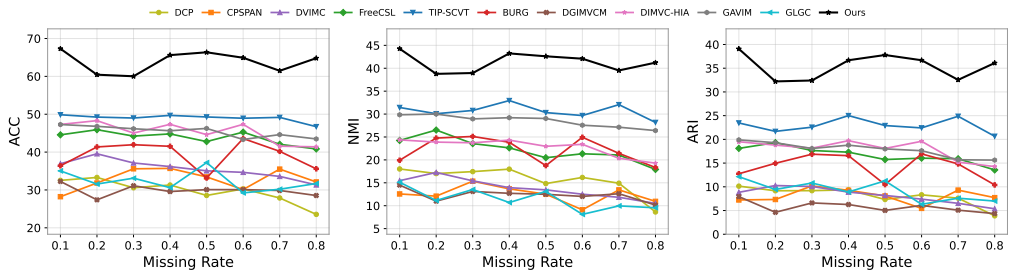


Figure 6: ACC, NMI, and ARI on Reuters-1200 under different requested missing settings.

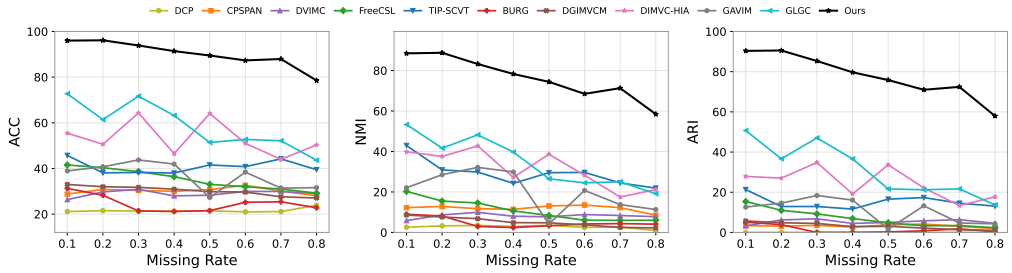


Figure 7: ACC, NMI, and ARI on NGs under different requested missing settings.

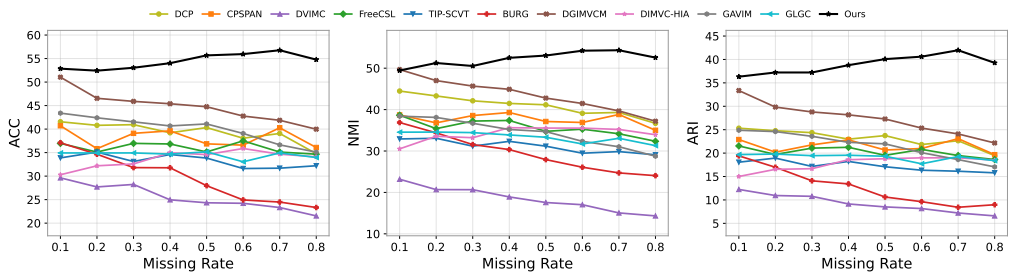


Figure 8: ACC, NMI, and ARI on Scene-15 under different requested missing settings.

Across all four datasets, TCRIMVC remains the best-performing method at the largest tested setting $\eta = 0.8$. On MNIST-10k, it exceeds the strongest baseline by 8.48 and 9.21 percentage points in NMI and ARI, complementing the 4.77-point ACC gain reported in the main paper. Its ACC/NMI/ARI improvements at $\eta = 0.8$ are $18.09/13.05/15.44$ points on Reuters-1200, $39.07/36.56/44.83$ points on NGs, and $18.69/15.88/19.61$ points on Scene-15. The advantage therefore persists across visual and textual data rather than being specific to MNIST-10k. The curves also show that TCRIMVC is considerably less sensitive to increasing incompleteness than several prediction- and graph-based baselines, supporting the use of direct observed evidence for recovery.

B Cross-view Image Completion Details

We instantiate the retrieval module in the latent spaces of frozen view-specific image autoencoders on MM-CelebA-HQ. After pairing and preprocessing, the dataset contains 29,996 RGB–sketch pairs resized to 128×128. Each target direction contains 8,999 target-missing test samples. The two frozen autoencoders produce 512-dimensional latent features and are trained for 100 epochs using Adam, a learning rate of 10^{-4}, a batch size of 64, and linear learning-rate decay.

TCRIMVC uses 10,198 complete pairs for training and 1,800 complete pairs for validation, a retrieval bank of 4,096 samples, $K = 20$, and $\tau = 0.1$. TCRIMVC and DCP (Lin et al., 2022) are trained for the same 200-epoch budget with identical masks and evaluation samples. Table 6 reports the complete quantitative results. TCRIMVC improves all three metrics in both directions, with the larger gain observed for Sketch→RGB.

Table 6: Cross-view image completion on MM-CelebA-HQ at $\eta = 0.3$. Lower RMSE/NMSE and higher PSNR are better.

Task	Method	RMSE ↓	NMSE ↓	PSNR ↑
Sketch→RGB	DCP	0.277912	0.930964	11.121849
	TCRIMVC	**0.242841**	**0.710826**	**12.293547**
RGB→Sketch	DCP	0.295050	1.330009	10.602079
	TCRIMVC	**0.280282**	**1.200191**	**11.048103**

Figures 9 and 10 present the selected qualitative examples. Each figure contains ten samples in a fixed horizontal order. The four rows show the observed source, held-out ground truth, DCP result, and TCRIMVC result.



Figure 9: Selected Sketch→RGB completion examples. Each row contains ten examples shown in the same horizontal order.

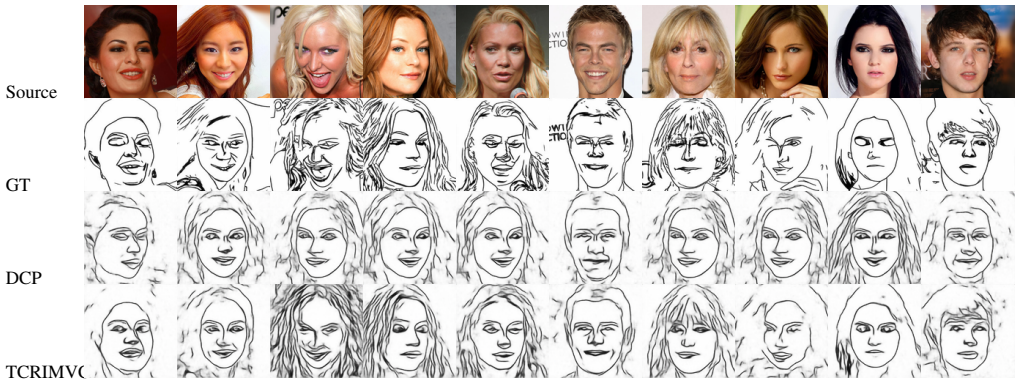


Figure 10: Selected RGB→Sketch completion examples. Each row contains ten examples shown in the same horizontal order.

The selected examples cover different identities, poses, hairstyles, and local facial structures. Compared with DCP, TCRIMVC generally produces target views with clearer identity-relevant structure and more consistent local details. Together with the quantitative gains, the results suggest that the retrieval mechanism captures transferable cross-view correspondence beyond the clustering objective.

C Properties of Direct-Context Retrieval

The following propositions characterize the evidence allocation and aggregation implemented in Eq. equation 5. They describe properties of the retrieval operation rather than generalization or performance guarantees.

Proposition 1 (Entropy-regularized evidence allocation)**.** *For a fixed Top-k set with scores $\{s_t\}_{t=1}^k$, where $k \geq 2$, the softmax weights*

$$w_t = \frac{\exp(s_t)}{\sum_{r=1}^k \exp(s_r)} \tag{16}$$

are the unique solution to

$$\max_{\mathbf{w} \in \Delta^{k-1}} \left\{ \sum_{t=1}^k w_t s_t + \mathcal{H}(\mathbf{w}) \right\}, \qquad \mathcal{H}(\mathbf{w}) = -\sum_{t=1}^k w_t \log w_t, \tag{17}$$

where $\Delta^{k-1} = \{\mathbf{w} \in \mathbb{R}_+^k : \sum_{t=1}^k w_t = 1\}$. For $k = 1$, the result reduces trivially to $w_1 = 1$.

Proof. For $k \geq 2$, define

$$J(\mathbf{w}) = \sum_{t=1}^{k} w_t s_t - \sum_{t=1}^{k} w_t \log w_t.$$

The first term is linear, while the entropy term is strictly concave on the probability simplex. Moreover, its one-sided derivative diverges at the boundary, so the maximizer lies in the relative interior. Introducing a Lagrange multiplier λ for $\sum_{t=1}^{k} w_t = 1$ gives

$$\frac{\partial}{\partial w_t}\left[J(\mathbf{w}) + \lambda\left(\sum_{r=1}^{k} w_r - 1\right)\right] = s_t - (1 + \log w_t) + \lambda = 0. \tag{18}$$

Hence, $w_t = \exp(s_t + \lambda - 1)$. Enforcing $\sum_{t=1}^{k} w_t = 1$ yields

$$w_t = \frac{\exp(s_t)}{\sum_{r=1}^{k} \exp(s_r)}.$$

Strict concavity of J establishes uniqueness. □

Proposition 2 (Convex-hull property of retrieved context)**.** *Let $\{(j_t, u_t)\}_{t=1}^{k}$ denote the selected evidence entries and assume $\|\mathbf{z}_{j_t}^{(u_t)}\|_2 \leq 1$. Define the pre-normalized retrieved context as*

$$\mathbf{r}_i^{(v)} = \sum_{t=1}^{k} w_{it}^{(v)} \mathbf{z}_{j_t}^{(u_t)}. \tag{19}$$

Then

$$\mathbf{r}_i^{(v)} \in \operatorname{conv}\left\{\mathbf{z}_{j_t}^{(u_t)}\right\}_{t=1}^{k}, \qquad \|\mathbf{r}_i^{(v)}\|_2 \leq 1. \tag{20}$$

If $\|\mathbf{r}_i^{(v)}\|_2 \geq \varepsilon$, the recovered representation $\widetilde{\mathbf{z}}_i^{(v)} = \nu(\mathbf{r}_i^{(v)})$ is the radial projection of $\mathbf{r}_i^{(v)}$ onto the unit sphere.

Proof. The softmax weights satisfy $w_{it}^{(v)} > 0$ and $\sum_{t=1}^{k} w_{it}^{(v)} = 1$. Therefore, $\mathbf{r}_i^{(v)}$ is a convex combination of the selected observed representations and consequently belongs to their convex hull. By the triangle inequality,

$$\begin{aligned} \|\mathbf{r}_i^{(v)}\|_2 &\leq \sum_{t=1}^{k} w_{it}^{(v)} \|\mathbf{z}_{j_t}^{(u_t)}\|_2 \\ &\leq \sum_{t=1}^{k} w_{it}^{(v)} = 1. \end{aligned} \tag{21}$$

When $\|\mathbf{r}_i^{(v)}\|_2 \geq \varepsilon$, the normalization operator satisfies

$$\nu(\mathbf{r}_i^{(v)}) = \frac{\mathbf{r}_i^{(v)}}{\|\mathbf{r}_i^{(v)}\|_2},$$

which is its radial projection onto the unit sphere. When the norm is smaller than ε, the stabilized operator instead returns $\mathbf{r}_i^{(v)}/\varepsilon$; in particular, a zero context remains zero. Since normalization may rescale the context by a factor greater than one, convex-hull membership is asserted only for the pre-normalized context. □