# PSEUDO-CORRESPONDENCE GUIDED BLIND-ALIGNMENT LEARNING FOR COMPLETELY UNALIGNED MULTI-VIEW CLUSTERING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multi-view clustering aims to exploit complementary and consistent information from heterogeneous views to discover a shared clustering structure without supervision. However, when cross-view instance correspondences are entirely unavailable, completely unaligned multi-view clustering faces three key challenges: unreliable local structures may hinder early correspondence discovery; correspondence estimation and representation learning are mutually dependent; and exact instance matching alone is insufficient to capture cluster-level semantics. To address these challenges, we propose Pseudo-Correspondence Guided Blind-Alignment Learning (PCGBAL), which treats cross-view alignment as an evolving latent-supervision process coupled with representation learning. PCGBAL first structurally bootstraps view-specific representations through masked topological reconstruction and then progressively discovers high-confidence pseudo-correspondences from the evolving representation space. These associations are fed back as latent supervision to refine heterogeneous representations, forming a mutually reinforcing correspondence–representation process. To further bridge the gap between instance-level correspondence and clustering semantics, instance-to-cluster contrastive alignment transfers cluster-level information across views beyond exact instance matches. Extensive experiments on eight benchmark datasets against seven competing methods demonstrate the effectiveness and robustness of PCGBAL in completely unaligned settings.

## 1 INTRODUCTION

Multi-view data describe the same underlying objects through heterogeneous sources, such as different sensors, feature extractors, or modalities. Multi-view clustering (MVC) aims to exploit complementary and consistent information across these views to discover a shared clustering structure in an unsupervised manner (Liang et al., 2025; Liu et al., 2026; Zhang et al., 2025a). Conventional MVC typically assumes fully known cross-view instance correspondences, where samples sharing the same index across different views refer to the same underlying object. However, this assumption can be difficult to satisfy in real-world scenarios, where cross-view correspondences may be disrupted by asynchronous data acquisition, communication delays, independently collected observations, or unavailable sample identifiers. When such correspondence information is absent, directly integrating samples according to their observed indices may incorrectly associate unrelated instances across views and consequently distort the underlying clustering structure (Li et al., 2026a; Sun et al., 2025).

Partially view-aligned clustering (PVC) relaxes this assumption by requiring only a subset of cross-view correspondences to be known. Existing PVC approaches typically exploit these aligned samples as correspondence priors to infer relationships among the remaining unaligned instances while simultaneously learning discriminative representations or consistent clustering structures (Zhou et al., 2026; Xu et al., 2026; Zhao et al., 2026b). Nevertheless, their alignment process still fundamentally depends on pre-given matched instances. Once all correspondence information becomes unavailable, these explicit alignment anchors disappear, leading to the more challenging setting of completely unaligned multi-view clustering. In this setting, both clustering structures and latent



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.

cross-view associations must be inferred directly from heterogeneous observations (Wen et al., 2024; Dong et al., 2026b; Dai et al., 2026).

Beyond the absence of cross-view correspondences, completely unaligned multi-view clustering is further complicated by the quality of the representation space from which correspondences are inferred. Local neighborhood structures constructed from raw observations may contain noisy or unreliable relations, which can distort early representation learning and consequently induce erroneous cross-view associations. More importantly, correspondence discovery and representation learning are intrinsically interdependent: reliable matching requires discriminative representations, whereas cross-view-consistent representations in turn benefit from trustworthy associations. Therefore, treating alignment as an isolated preprocessing step or estimating correspondences from a fixed representation space may prevent the two processes from mutually improving each other. Moreover, exact instance correspondences provide only identity-level supervision and cannot fully characterize the semantic relationships among samples belonging to the same latent cluster.

These observations motivate us to couple blind alignment with progressive representation learning. Accordingly, we propose Pseudo-Correspondence Guided Blind-Alignment Learning (PCGBAL), which progressively discovers reliable cross-view associations and uses them as latent supervision to refine representations, while incorporating cluster-level semantics beyond instance matching. The overall framework is illustrated in Figure 1. The main contributions are summarized as follows:

- We introduce a structural bootstrapping strategy based on masked topological reconstruction to reduce the influence of unreliable local relations. By recovering masked graph connections from the remaining topology, the proposed strategy encourages more stable and structurally robust view-specific representations.
- We formulate correspondence discovery and representation learning as a progressive coupling process rather than two isolated stages. Pseudo-correspondences are progressively identified from evolving representations and fed back to refine cross-view representations, thereby improving the reliability of subsequent correspondence estimation.
- We extend cross-view alignment beyond exact instance matching by incorporating cluster-level semantic information. Prototype-guided instance-to-cluster contrastive learning transfers clustering semantics across views and promotes semantic consistency among heterogeneous representations.
- Extensive experiments on eight benchmark datasets against seven competing methods demonstrate the effectiveness and robustness of PCGBAL. Additional ablation, statistical, and robustness analyses further validate the proposed framework under completely unaligned settings.

## 2 Related Work

### 2.1 Deep Multi-View Clustering

Deep multi-view clustering has increasingly shifted from shallow feature integration toward nonlinear representation learning. Deep autoencoder-based approaches commonly learn compact view-specific or shared latent representations (Wang et al., 2024; Li et al., 2024a), while graph-based models further incorporate neighborhood structures and local geometric relationships into representation learning (Xiao et al., 2023; Yan et al., 2026). More recently, contrastive learning has become an effective mechanism for exploiting cross-view consistency while preserving discriminative information. For example, Xiao et al. (2025) construct a view graph to select informative view pairs and a cross-view sample graph to generate reliable sample weights for contrastive learning, thereby reducing the influence of low-quality views and unreliable samples. Fei et al. (2025) further combine graph convolution with similarity-guided instance-level and cluster-level contrastive learning, enabling structural information and clustering-oriented semantic consistency to be jointly exploited.

Recent studies have further incorporated more expressive architectures and semantic priors into multi-view representation learning. For example, Zhang et al. (2026) exploit Mamba-based sequence modeling to capture cross-view dependencies and global-local structural relationships among samples. Ouyang et al. (2026) construct category prototypes using pretrained vision-language models, while Zhang et al. (2025b) combine diffusion-based generation with contrastive

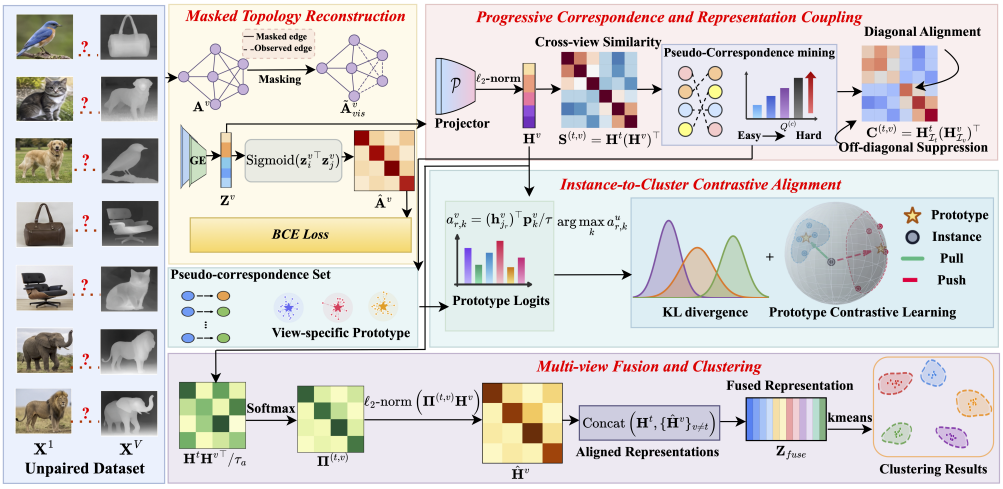


Figure 1: Overview of PCGBAL. Masked topological reconstruction structurally bootstraps the representation space. Evolving representations are then used to progressively discover pseudo-correspondences, which are fed back through cross-view feature synchronization to refine the representations, forming a correspondence–representation feedback process. Instance-to-cluster contrastive alignment further complements identity-level supervision with cluster-level semantics before multi-view fusion and final clustering.

learning for incomplete multi-view clustering. These developments illustrate the increasing capability of deep models to capture nonlinear and structural information from heterogeneous views.

Despite these advances, most deep multi-view representation learning methods operate under settings in which cross-view correspondence information is available or at least partially preserved. Such correspondence information enables direct cross-view comparison and provides reliable guidance for learning consistent representations. When cross-view correspondences are entirely unavailable, these alignment signals cannot be directly constructed, making both cross-view association discovery and representation learning substantially more challenging.

### 2.2 Unaligned Multi-View Clustering

Early studies on unaligned multi-view clustering mainly recover latent cross-view associations through permutation estimation, graph matching, or anchor-based alignment (Cao et al., 2024; Li et al., 2024b; Cao et al., 2026). These methods exploit transferable structural information across views to alleviate the absence of explicit sample correspondences.

More recent studies further explore probabilistic alignment and reliable structural guidance without requiring pre-given alignment priors. For example, Dong et al. (2026a) integrate permutation estimation into bipartite graph learning and derive cross-view correspondences through a probabilistic transition process with adaptive template selection. Xin et al. (2025) improve clustering reliability through multilevel clustering, synthesized-view alignment, and cross-view guidance. Xin et al. (2026) construct a synthesized view to reduce inter-view distribution discrepancies and transfer both global and local structural information across views. For large-scale scenarios, Li et al. (2026b) combine anchor alignment with bipartite graph learning and introduce a shifted graph learning strategy to preserve clustering information and graph energy.

Existing unaligned MVC methods mainly treat correspondence estimation as an alignment variable or a matching stage, whereas PCGBAL treats pseudo-correspondences as dynamically evolving latent supervision that is repeatedly regenerated from the representation space and fed back into both instance-level and cluster-level representation learning.

# 3 METHODOLOGY

This section first formulates the completely unaligned multi-view clustering problem together with its underlying motivation, followed by the proposed framework and its optimization objective.

## 3.1 PROBLEM FORMULATION AND MOTIVATION

Let $\{\mathbf{X}^v \in \mathbb{R}^{N\times d_v}\}_{v=1}^V$ denote a multi-view dataset with $N$ instances across $V$ views, where $d_v$ is the feature dimension of the $v$-th view. Under the completely unaligned setting, instance-level correspondences across views are entirely unknown. The objective is to jointly discover latent cross-view alignments and partition the $N$ instances into $K$ semantic clusters.

Completely unaligned multi-view clustering is challenging due to the inherent interdependence between correspondence discovery and representation learning. Reliable cross-view matching requires a discriminative representation space, yet such a space is difficult to obtain without trustworthy correspondence supervision. This difficulty is further amplified by noisy local structures in the early training stage and by the fact that exact instance matches alone cannot fully capture shared cluster semantics. These observations motivate a learning process that first stabilizes the representation space and then progressively improves cross-view consistency using increasingly reliable instance- and cluster-level supervision.

## 3.2 STRUCTURAL BOOTSTRAPPING VIA MASKED TOPOLOGICAL RECONSTRUCTION

Reliable cross-view correspondence discovery requires a sufficiently robust representation space. Constructing $k$-nearest neighbor (KNN) graphs is a common way to capture intrinsic geometric structures, but graphs obtained directly from raw observations may contain noisy or unreliable connections. If these relations are overfitted during early training, the resulting representations may produce erroneous cross-view matches. Inspired by the regularization effect of random dropout in neural networks, we introduce Masked Topological Manifold Modeling (MTMM), which randomly masks a subset of graph connections and reconstructs them from the remaining visible topology. Specifically, for the $v$-th view, we construct a symmetric $k$-NN graph with edge set $\mathcal{E}^v$. At each training epoch, we uniformly sample and remove a subset of edges $\mathcal{E}^v_{mask} \subset \mathcal{E}^v$ with a ratio of $p$. Message passing relies solely on the remaining visible edge set $\mathcal{E}^v_{vis} = \mathcal{E}^v \setminus \mathcal{E}^v_{mask}$, yielding the normalized adjacency matrix $\tilde{\mathbf{A}}^v_{vis}$. A Graph Convolutional Network (GCN) then propagates the fully visible features over this sparsified topology:

$$\mathbf{Z}^v = \ell_2\text{-norm}\Big(\sigma_L(\tilde{\mathbf{A}}^v_{vis}\sigma_L(\tilde{\mathbf{A}}^v_{vis}\mathbf{X}^v\mathbf{W}^v_1)\mathbf{W}^v_2)\Big), \tag{1}$$

where $\mathbf{W}^v_1$ and $\mathbf{W}^v_2$ are learnable weights, $\sigma_L(\cdot)$ denotes the LeakyReLU activation function, and $\ell_2$-norm$(\cdot)$ denotes row-wise normalization to stabilize the embedding scale.

Based on the obtained representations $\mathbf{Z}^v$, the masked connections are predicted as $\hat{A}^v_{ij} = \mathrm{Sigmoid}((\mathbf{z}^v_i)^\top \mathbf{z}^v_j)$. An equal number of disconnected node pairs are sampled as negative edges, i.e., $|\mathcal{E}^v_{neg}| = |\mathcal{E}^v_{mask}|$. We then adopt a balanced binary cross-entropy (BCE) loss (Ruby et al., 2020) over the masked positive edges and sampled negative edges:

$$\mathcal{L}_{rec} = \frac{1}{V}\sum_{v=1}^{V}\left(-\frac{1}{|\mathcal{E}^v_{mask}|}\sum_{(i,j)\in\mathcal{E}^v_{mask}}\log\hat{A}^v_{ij} - \frac{1}{|\mathcal{E}^v_{neg}|}\sum_{(i,j)\in\mathcal{E}^v_{neg}}\log(1-\hat{A}^v_{ij})\right). \tag{2}$$

The first term encourages the recovery of masked true connections, while the second term suppresses nonexistent links to prevent trivial positive predictions. By reconstructing the masked topology from representations learned over the visible graph $\tilde{\mathbf{A}}^v_{vis}$, the learned representation $\mathbf{Z}^v$ is encouraged to capture structural dependencies beyond individual local connections, providing a more reliable basis for subsequent correspondence discovery.

## 3.3 PROGRESSIVE CORRESPONDENCE AND REPRESENTATION COUPLING

The structurally robust representations obtained above provide the basis for progressively discovering reliable cross-view correspondences and using them to refine the representation space. Let

$\mathcal{P}$ denote a shared projection head and $\mathbf{H}^v = \ell_2\text{-norm}(\mathcal{P}(\mathbf{Z}^v))$ the projected embeddings. For the template view $t$ and an auxiliary view $v \neq t$, we compute the similarity matrix $\mathbf{S}^{(t,v)} = \mathbf{H}^t(\mathbf{H}^v)^\top$. For each sample, the top-$\kappa$ most similar samples in the opposite view are selected, denoted by $\mathcal{N}_\kappa^v(i)$ and $\mathcal{N}_\kappa^t(j)$, respectively. The reciprocal and relaxed candidate sets are defined as

$$\mathcal{M}^{(t,v)} = \{(i,j) \mid j \in \mathcal{N}_\kappa^v(i),\ i \in \mathcal{N}_\kappa^t(j)\}, \quad \mathcal{R}^{(t,v)} = \{(i,j) \mid j \in \mathcal{N}_\kappa^v(i) \ \vee \ i \in \mathcal{N}_\kappa^t(j)\}. \tag{3}$$

where $\vee$ denotes the logical OR operator. Thus, pairs in $\mathcal{M}^{(t,v)}$ satisfy the top-$\kappa$ neighborhood condition in both directions and are regarded as high-confidence reciprocal candidates, while $\mathcal{R}^{(t,v)}$ contains relaxed candidates that satisfy the condition in at least one direction. We rank the reciprocal candidates by similarity and greedily retain non-conflicting pairs under a one-to-one constraint. To improve correspondence coverage, a completion ratio $\eta \in (0,1]$ specifies the target number of correspondences as $\lceil \eta N \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function. If fewer reciprocal pairs are obtained, maximum-cardinality bipartite matching is performed over the remaining candidates in $\mathcal{R}^{(t,v)}$. The additional pairs are ranked by similarity and appended after the reciprocal pairs, yielding the ordered set $\bar{\mathcal{M}}_\eta^{(t,v)}$.

Since correspondences estimated from immature representations may be unreliable, we introduce Dynamic Self-Paced Pseudo-Correspondence Mining (DSPPM) to follow an easy-to-hard learning curriculum, prioritizing high-confidence pairs before gradually incorporating more challenging ones. At epoch $c \in \{0, \ldots, C-1\}$, the retention ratio and number of selected correspondences are

$$\rho^{(c)} = \rho_{\min} + \frac{c}{C-1}(\rho_{\max} - \rho_{\min}), \qquad Q^{(c)} = \max\left(1, \left\lceil \rho^{(c)} |\bar{\mathcal{M}}_\eta^{(t,v)}| \right\rceil\right), \tag{4}$$

and the retained pseudo-correspondence set is $\mathcal{A}^{(c)} = \bar{\mathcal{M}}_\eta^{(t,v)}[1 : Q^{(c)}]$. Thus, training begins with a small set of reliable and relatively easy correspondences and progressively incorporates more challenging pairs as the representation space improves.

Given the retained correspondences, we further introduce Cross-View Feature Synchronization (CFS) to refine heterogeneous representations. Let $\mathcal{A}^{(c)} = \{(i_r, j_r)\}_{r=1}^{N_a^{(c)}}$, where $N_a^{(c)}$ denotes the number of selected pairs. With $\mathcal{I}_t = \{i_r\}$ and $\mathcal{I}_v = \{j_r\}$, their cross-view correlation matrix is $\mathbf{C}^{(t,v)} = \mathbf{H}_{\mathcal{I}_t}^t (\mathbf{H}_{\mathcal{I}_v}^v)^\top$. The diagonal entries correspond to selected pseudo-correspondences, while the off-diagonal entries measure correlations between different pairs. CFS encourages strong correlations for matched pairs while suppressing excessive off-diagonal correlations:

$$\mathcal{L}_{fs}^{(t,v)} = \frac{1}{N_a^{(c)}} \sum_{r=1}^{N_a^{(c)}} \max\{0, m - C_{rr}^{(t,v)}\}^2 + \alpha \frac{1}{N_a^{(c)}(N_a^{(c)}-1)} \sum_{r \neq s} (C_{rs}^{(t,v)})^2, \tag{5}$$

where $m$ denotes the positive correlation margin and $\alpha$ controls the off-diagonal regularization strength. The overall synchronization objective is

$$\mathcal{L}_{fs} = \frac{1}{V-1} \sum_{v \neq t} \mathcal{L}_{fs}^{(t,v)}. \tag{6}$$

### 3.4 Instance-to-Cluster Contrastive Alignment

The correspondence–representation process above primarily operates at the instance level: a selected pair $(i_r, j_r)$ indicates that two observations are likely to describe the same underlying object. However, correspondence identity and clustering semantics are not equivalent. Two samples without an exact correspondence may still belong to the same latent cluster. Consequently, relying only on instance-level synchronization may introduce undesirable semantic false negatives. To complement identity-level alignment with cluster-level consistency, we introduce Instance-to-Cluster Contrastive Alignment (I2C-CA), which transfers cluster-level information across views through view-specific prototypes.

For each view $u \in \{t, v\}$, we introduce a trainable prototype bank $\mathbf{P}^u = [\mathbf{p}_1^u, \ldots, \mathbf{p}_K^u]^\top \in \mathbb{R}^{K \times d}$, where each prototype represents a latent cluster. Both the instance embeddings and cluster prototypes are $\ell_2$-normalized. Given the selected pseudo-correspondence set $\mathcal{A}^{(c)} = \{(i_r, j_r)\}_{r=1}^{N_a^{(c)}}$,

the prototype logits are computed as $a^t_{r,k} = (\mathbf{h}^t_{i_r})^\top \mathbf{p}^t_k/\tau$ and $a^v_{r,k} = (\mathbf{h}^v_{j_r})^\top \mathbf{p}^v_k/\tau$, where $a^u_{r,k}$ denotes the temperature-scaled similarity between the $r$-th selected instance and the $k$-th prototype in view $u$, and $\tau$ is the contrastive temperature. The corresponding hard cluster assignments are $y^t_r = \arg\max_k a^t_{r,k}$ and $y^v_r = \arg\max_k a^v_{r,k}$. For each selected pair, the cluster assignment predicted by one view determines the positive prototype in the other view, while the remaining prototypes serve as negatives. The two views therefore provide symmetric cross-view supervision, with gradients stopped through the target assignments. The prototype-based contrastive objective is defined as

$$\mathcal{L}^{(t,v)}_{con} = -\frac{1}{2N^{(c)}_a} \sum_{r=1}^{N^{(c)}_a} \left[ \log \frac{\exp(a^v_{r,y^t_r})}{\sum_{k=1}^K \exp(a^v_{r,k})} + \log \frac{\exp(a^t_{r,y^v_r})}{\sum_{k=1}^K \exp(a^t_{r,k})} \right]. \tag{7}$$

To avoid degenerate assignments in which most selected instances collapse onto a few prototypes, a cluster-balance regularizer is further introduced. Let $\mathbf{q}^u_r = \mathrm{Softmax}(\mathbf{a}^u_r)$ denote the cluster probability distribution and $\bar{\mathbf{q}}^u = (N^{(c)}_a)^{-1} \sum_{r=1}^{N^{(c)}_a} \mathbf{q}^u_r$ denote its average distribution. The balance regularizer is defined as

$$\mathcal{L}^{(t,v)}_{bal} = \frac{1}{2} \sum_{u \in \{t,v\}} \mathrm{KL}\left(\bar{\mathbf{q}}^u \,\|\, \mathbf{u}_K\right), \tag{8}$$

where $\mathbf{u}_K$ denotes the uniform distribution over $K$ clusters. The instance-to-cluster contrastive alignment loss between the template view and the auxiliary view $v$ is

$$\mathcal{L}^{(t,v)}_{i2c} = \mathcal{L}^{(t,v)}_{con} + \beta \mathcal{L}^{(t,v)}_{bal}, \tag{9}$$

where $\beta$ controls the cluster-balance regularization. Finally, the overall objective is

$$\mathcal{L}_{i2c} = \frac{1}{V-1} \sum_{v \neq t} \mathcal{L}^{(t,v)}_{i2c}. \tag{10}$$

Thus, PCGBAL performs complementary alignment at two levels: instance-level pseudo-correspondences provide identity-specific cross-view supervision, while prototype-based alignment preserves cluster-level semantic consistency beyond exact instance matching.

### 3.5 Overall Objective

The above objectives jointly contribute to progressive blind alignment. $\mathcal{L}_{rec}$ structurally bootstraps the representation space before reliable correspondences are available; $\mathcal{L}_{fs}$ converts progressively discovered instance associations into cross-view representation supervision; and $\mathcal{L}_{i2c}$ complements identity-level synchronization with cluster-level semantic consistency. They jointly instantiate the correspondence–representation interaction described above, leading to the overall objective

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{fs} + \lambda_2 \mathcal{L}_{i2c}, \tag{11}$$

where $\lambda_1$ and $\lambda_2$ are trade-off hyperparameters.

### 3.6 Multi-View Fusion and Clustering

After training, we perform cross-view fusion. For each auxiliary view $v \neq t$, we construct a soft correspondence matrix $\mathbf{\Pi}^{(t,v)} \in \mathbb{R}^{N \times N}$ as

$$\mathbf{\Pi}^{(t,v)} = \mathrm{Softmax}_{row}\left(\mathbf{H}^t (\mathbf{H}^v)^\top / \tau_a\right), \qquad \hat{\mathbf{H}}^v = \ell_2\text{-norm}\left(\mathbf{\Pi}^{(t,v)} \mathbf{H}^v\right), \tag{12}$$

where $\tau_a$ is the alignment temperature and $\Pi^{(t,v)}_{ij}$ represents the soft correspondence weight between the $i$-th sample in the template view and the $j$-th sample in view $v$. The row-wise Softmax ensures that each template sample aggregates auxiliary-view features according to their cross-view similarities. Accordingly, $\mathbf{\Pi}^{(t,v)} \mathbf{H}^v$ maps the auxiliary-view representations into the template-view sample order, and the resulting aligned representation $\hat{\mathbf{H}}^v$ is further $\ell_2$-normalized to maintain a consistent feature scale across views.

The final fused representation is obtained by concatenating the template representation with all softly aligned auxiliary representations as $\mathbf{Z}_{fuse} = \mathrm{Concat}\left(\mathbf{H}^t, \{\hat{\mathbf{H}}^v\}_{v\neq t}\right)$, where $\mathrm{Concat}(\cdot)$ denotes feature-wise concatenation. $\mathbf{Z}_{fuse}$ is then fed into $k$-means to obtain the final clustering assignments. The complete pseudocode of PCGBAL is provided in Algorithm 1 in Appendix A.1. A detailed computational complexity analysis is provided in Appendix A.3.

# 4 EXPERIMENTS

The experiments are designed to evaluate the proposed framework by addressing the following research questions: **Q1:** Does our method outperform state-of-the-art multi-view clustering methods in the completely unaligned setting? **Q2:** How does each proposed component affect the overall performance? **Q3:** How robust is the proposed method across different unaligned rates? Due to space constraints, the remaining questions are thoroughly investigated in the **Appendix**: **Q4:** Are the performance improvements statistically consistent across datasets? **Q5:** Do the learned representations exhibit class-separable structure in a low-dimensional visualization? **Q6:** How sensitive is the model to the choice of template view and other hyperparameters? **Q7:** Does the model exhibit stable convergence during optimization? **Q8:** How does the running time of the proposed framework compare with that of existing methods? **Q9:** How robust is the model against noise?

## 4.1 DATASETS AND BASELINES

Experiments are conducted on eight multi-view benchmark datasets: **Handwritten** (Wen et al., 2024), **Scene15** (Fei-Fei & Perona, 2005), **MNIST-USPS** (Peng et al., 2019), **GSE** (Deng et al., 2025), **Hdigit** (Wen et al., 2024), **ALOI** (Geusebroek et al., 2005), **ReutersEF** (Huang et al., 2019), and **Cifar10** (Zhang et al., 2019). Dataset statistics are provided in **Appendix** A.2. Following standard protocols, the unaligned rate is defined as $(N - N_a)/N$, where $N_a$ denotes the number of preserved cross-view correspondences. The completely unaligned setting corresponds to $N_a = 0$.

We compare against seven state-of-the-art methods for partially aligned or completely unaligned multi-view clustering. **DAGF** (Jiang et al., 2026) and **FUMC** (Li et al., 2024b) are shallow methods, while **BIN** (Deng et al., 2026), **VITAL** (He et al., 2024), **CANDY** (Guo et al., 2024), **KNNDA** (Zhao et al., 2026a), and **AFUMC** (Sun et al., 2026) are deep methods.

## 4.2 IMPLEMENTATION DETAILS

All experiments are conducted on Ubuntu with an AMD EPYC 7513 CPU, 256 GB RAM, and NVIDIA RTX 4090 GPUs. Our method and the Python-based baselines are implemented using Python 3.9 with PyTorch, while the MATLAB-based baselines are evaluated using MATLAB R2025b. All competing methods use their official implementations, and their hyperparameters are tuned by grid search within the ranges recommended in the original papers. Each experiment is repeated 10 times, and the average results are reported. Performance is evaluated by ACC, NMI, Purity, F-score, Precision, and ARI, all reported in percentages.

For the proposed model, the first view is designated as the template view in all experiments. We construct $k$-NN graphs with $k = 5$ and employ a graph autoencoder with a 2000-dimensional hidden layer and a 100-dimensional embedding. The model is trained for 200 epochs with a batch size of 2000 using Adam (Kingma & Ba, 2014), with a learning rate of $10^{-4}$ and weight decay of $0.01$. We set $m = 0.95$, $\alpha = 0.05$, $\tau = 0.5$, $\beta = 0.1$, $\tau_a = 0.1$, $\rho_{\min} = 0.1$, $\rho_{\max} = 1.0$, and $\eta = 0.8$. The searched parameters are $\lambda_1, \lambda_2 \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, $p \in \{0.2, 0.3, 0.4\}$, and $\kappa \in \{10, 20, 30, 50\}$.

## 4.3 EXPERIMENTAL RESULTS AND ANALYSIS (**Q1**)

The clustering results on eight benchmark datasets under the completely unaligned setting are reported in Tables 1 and 2. Overall, our method achieves consistently strong performance across most datasets and evaluation metrics, demonstrating the effectiveness of progressive pseudo-correspondence mining and cross-view semantic alignment when no cross-view correspondence information is available. On the large-scale Cifar10 dataset, our method also maintains competi-

Table 1: Clustering performance on Handwritten, Scene15, MNIST-USPS, and GSE. The best and second-best results are shown in bold and underlined, respectively. 'Pur', 'Fsc', and 'Pre' denote Purity, F-score, and Precision, respectively.

| Method | Handwritten | | | | | | Scene15 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | NMI | Pur | Fsc | Pre | ARI | ACC | NMI | Pur | Fsc | Pre | ARI |
| DAGF | 64.50 | 63.75 | 64.50 | 54.49 | 50.59 | 49.03 | 45.95 | 41.30 | 46.06 | 33.45 | 25.55 | 26.83 |
| FUMC | 14.05 | 1.17 | 14.40 | 10.38 | 10.07 | 0.13 | 32.17 | 34.05 | 34.78 | 25.49 | 23.25 | 19.39 |
| BIN | 81.15 | 69.48 | 81.25 | 68.06 | 67.70 | 64.05 | 19.38 | 20.36 | 25.84 | 15.01 | 15.75 | 7.76 |
| VITAL | 68.05 | 55.46 | 68.05 | 67.11 | 68.78 | 46.09 | 39.40 | 37.81 | 42.43 | 38.05 | 38.93 | 23.76 |
| CANDY | 67.75 | 57.79 | 67.75 | 67.01 | 67.72 | 48.02 | 17.15 | 11.55 | 19.78 | 17.36 | 19.88 | 5.22 |
| KNNDA | 43.35 | 41.24 | 46.05 | 42.66 | 44.19 | 30.05 | 24.50 | 23.83 | 27.49 | 24.17 | 25.78 | 10.51 |
| AFUMC | 85.20 | 75.04 | 85.20 | 74.21 | 74.16 | 70.94 | 43.23 | 42.23 | 46.58 | 31.79 | 32.24 | 25.75 |
| Ours | **91.45** | **84.26** | **91.45** | **91.48** | **91.80** | **82.47** | **47.76** | **46.67** | **48.92** | **43.57** | **43.63** | **30.54** |
| Method | MNIST-USPS | | | | | | GSE | | | | | |
| | ACC | NMI | Pur | Fsc | Pre | ARI | ACC | NMI | Pur | Fsc | Pre | ARI |
| DAGF | 44.46 | 39.65 | 45.56 | 34.45 | 28.05 | 25.29 | 62.09 | 70.32 | 63.83 | 56.21 | 46.77 | 53.97 |
| FUMC | 38.90 | 33.75 | 42.16 | 29.87 | 22.05 | 18.91 | 35.39 | 47.32 | 37.54 | 31.30 | 24.43 | 27.54 |
| BIN | 26.92 | 12.75 | 30.56 | 17.53 | 17.63 | 7.87 | 71.20 | 76.43 | 74.96 | 62.68 | 62.71 | 62.29 |
| VITAL | 49.44 | 38.53 | 53.02 | 48.74 | 49.73 | 28.99 | 69.60 | 72.16 | 71.89 | 65.22 | 66.30 | 59.57 |
| CANDY | 45.10 | 40.93 | 50.42 | 45.24 | 47.59 | 27.52 | 69.45 | 73.58 | 74.37 | 65.61 | 68.22 | 60.44 |
| KNNDA | 21.82 | 12.92 | 25.22 | 22.50 | 25.99 | 5.90 | 35.11 | 40.84 | 38.22 | 32.23 | 34.35 | 22.89 |
| AFUMC | 62.38 | 56.36 | 65.88 | 52.44 | 51.64 | 45.00 | 65.22 | 72.39 | 68.99 | 57.59 | 57.08 | 53.30 |
| Ours | **73.48** | **67.93** | **73.48** | **73.33** | **76.70** | **57.74** | **73.40** | **77.95** | **76.22** | **68.37** | **70.74** | **64.68** |

Table 2: Clustering performance on Hdigit, ALOI, ReutersEF, and Cifar10. The best and second-best results are shown in bold and underlined, respectively. 'Pur', 'Fsc', and 'Pre' denote Purity, F-score, and Precision, respectively.

| Method | Hdigit | | | | | | ALOI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | NMI | Pur | Fsc | Pre | ARI | ACC | NMI | Pur | Fsc | Pre | ARI |
| DAGF | 47.61 | 44.06 | 51.44 | 37.34 | 29.61 | 28.30 | 36.52 | 47.12 | 37.95 | 5.62 | 2.98 | 3.82 |
| FUMC | 28.23 | 15.87 | 29.15 | 20.38 | 18.78 | 10.69 | 51.85 | 68.64 | 55.14 | 38.62 | 35.02 | 37.94 |
| BIN | 32.45 | 18.53 | 37.29 | 21.93 | 22.14 | 12.82 | 49.48 | 71.71 | 54.16 | 43.37 | 41.81 | 37.36 |
| VITAL | 58.87 | 44.27 | 58.87 | 59.70 | 61.59 | 34.63 | 37.44 | 59.02 | 39.22 | 35.69 | 36.19 | 25.51 |
| CANDY | 36.35 | 32.37 | 42.20 | 37.36 | 43.91 | 17.82 | 27.38 | 58.71 | 30.92 | 25.43 | 27.86 | 19.56 |
| KNNDA | 35.85 | 29.62 | 39.64 | 34.71 | 35.25 | 17.72 | 32.59 | 59.13 | 35.74 | 30.12 | 32.04 | 23.20 |
| AFUMC | 61.49 | 53.97 | 64.65 | 50.65 | 49.40 | 43.41 | 53.19 | **74.92** | **56.65** | 50.43 | 46.01 | 39.44 |
| Ours | **74.61** | **67.76** | **74.61** | **74.28** | **75.98** | **59.48** | **54.13** | 69.72 | 55.73 | **52.90** | **53.99** | **40.93** |
| Method | ReutersEF | | | | | | Cifar10 | | | | | |
| | ACC | NMI | Pur | Fsc | Pre | ARI | ACC | NMI | Pur | Fsc | Pre | ARI |
| DAGF | 28.13 | 2.74 | 28.84 | 34.65 | 21.29 | -0.27 | NaN | NaN | NaN | NaN | NaN | NaN |
| FUMC | 45.15 | 28.67 | 55.38 | 36.76 | 40.87 | 21.68 | 88.12 | 78.02 | 88.12 | 78.98 | 78.45 | 76.63 |
| BIN | 28.53 | 4.94 | 42.34 | 22.46 | 24.89 | 4.13 | NaN | NaN | NaN | NaN | NaN | NaN |
| VITAL | 42.33 | 24.45 | 50.98 | 35.69 | 42.48 | 19.32 | 88.00 | 75.64 | 88.00 | 87.98 | 88.08 | 75.67 |
| CANDY | 23.72 | 3.92 | 32.50 | 16.72 | 17.16 | 0.14 | 60.81 | 54.84 | 60.81 | 65.31 | 83.19 | 26.10 |
| KNNDA | 47.29 | 24.83 | 53.73 | 29.23 | 31.43 | 16.00 | NaN | NaN | NaN | NaN | NaN | NaN |
| AFUMC | 42.14 | 23.39 | 52.02 | 36.04 | 38.92 | 17.38 | 83.37 | 68.55 | 83.37 | 70.75 | 70.80 | 67.06 |
| Ours | **65.87** | **43.54** | **68.21** | **61.64** | **63.04** | **37.96** | **89.13** | **78.63** | **89.13** | **89.33** | **90.02** | **77.33** |

tive performance, while DAGF, BIN, and KNNDA do not finish within the 12-hour time limit and are therefore reported as NaN. On ALOI, although AFUMC achieves higher NMI and Purity, our method obtains the best results on ACC, F-score, Precision, and ARI.

Table 3: Ablation results on Handwritten and Scene15. The best results are shown in bold. 'Pur', 'Fsc', and 'Pre' denote Purity, F-score, and Precision, respectively.

| Variant | Handwritten | | | | | | Scene15 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | NMI | Pur | Fsc | Pre | ARI | ACC | NMI | Pur | Fsc | Pre | ARI |
| w/o MTMM | 87.09 | 80.25 | 87.09 | 87.13 | 88.27 | 75.21 | 27.12 | 29.71 | 32.82 | 26.56 | 29.04 | 13.82 |
| w/o SPS | 89.33 | 82.48 | 89.33 | 89.36 | 89.94 | 79.09 | 43.18 | 45.12 | 46.48 | 39.66 | 40.75 | 29.59 |
| w/o MP | 84.50 | 78.56 | 84.50 | 84.83 | 85.71 | 71.95 | 42.21 | 44.79 | 45.17 | 39.37 | 40.77 | 29.07 |
| w/o CFS | 88.96 | 81.93 | 88.96 | 88.95 | 89.69 | 78.25 | 43.50 | 45.77 | 48.07 | 40.49 | 42.43 | 29.49 |
| w/o I2C-CA | 88.87 | 81.90 | 88.87 | 88.91 | 89.73 | 78.15 | 43.09 | 44.14 | 46.16 | 39.66 | 40.87 | 29.00 |
| **Full** | **91.45** | **84.26** | **91.45** | **91.48** | **91.80** | **82.47** | **47.76** | **46.67** | **48.92** | **43.57** | **43.63** | **30.54** |

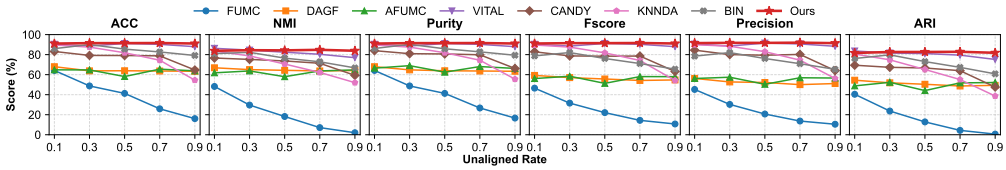


Figure 2: Clustering performance on Handwritten under different unaligned rates.

### 4.4 Ablation Study (Q2)

To investigate the contributions of the core components, we construct five degraded variants: **w/o MTMM** removes topology masking while retaining graph reconstruction; **w/o SPS** removes self-paced selection and uses the full pseudo-correspondence set throughout training; **w/o MP** replaces the mined pairs with an equal number of randomly sampled, collision-free one-to-one cross-view pairs; **w/o CFS** disables cross-view feature synchronization by setting $\lambda_1 = 0$; and **w/o I2C-CA** disables instance-to-cluster contrastive alignment by setting $\lambda_2 = 0$.

Table 3 reports the results on two datasets, while those on the remaining six datasets are provided in **Appendix** B. The full model consistently outperforms all five degraded variants. MTMM provides a robust structural basis, SPS controls the reliability of pseudo-correspondence supervision, CFS refines cross-view representations, and I2C-CA promotes cluster-level semantic consistency. Moreover, the matched-budget w/o MP control shows that the performance gains arise from the mined correspondences rather than merely from introducing cross-view pairwise supervision.

### 4.5 Robustness to Different Unaligned Rates (Q3)

We evaluate robustness to correspondence corruption on Handwritten by varying the unaligned rate over $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. As shown in Figure 2, our method maintains consistently strong performance across all six metrics as the unaligned rate increases, demonstrating stable clustering behavior under different degrees of cross-view misalignment. In contrast, most competing methods exhibit noticeable performance degradation at higher unaligned rates. The results indicate that progressively mining reliable pseudo-correspondences enables the proposed framework to remain robust even when cross-view correspondence information becomes increasingly unreliable.

## 5 Conclusion

This paper proposes Pseudo-Correspondence Guided Blind-Alignment Learning (PCGBAL) for completely unaligned multi-view clustering. PCGBAL first learns robust structural representations through masked topological reconstruction and then progressively couples pseudo-correspondence discovery with cross-view representation refinement. Instance-to-cluster contrastive alignment further introduces cluster-level semantic supervision beyond exact instance matching. Extensive experiments on eight benchmark datasets demonstrate the effectiveness and robustness of the proposed framework under completely unaligned settings. Future work will extend PCGBAL to incomplete and unaligned multi-view clustering, where both view observations and cross-view correspondences are partially unavailable.

### AI USE STATEMENT

In this work, we used generative AI tools (specifically ChatGPT) to improve the readability, grammar, and phrasing of the manuscript. We have not used generative AI tools to generate novel scientific ideas, design the methodology, or conduct experiments, and no data or results were fabricated using AI. We have reviewed all AI-assisted work. Specifically, the authors carefully proofread and manually verified all AI-polished sections to ensure they accurately reflect our original scientific intent, methodology, and findings without introducing any inaccurate or hallucinated claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our work, we have made the complete source code, preprocessed datasets, and detailed execution instructions publicly available via an anonymous repository: `https://anonymous.4open.science/r/PCGBAL-D72D/`.

## REFERENCES

Junfeng Cao, Wenhua Dong, and Jing Chen. View-unaligned clustering with graph regularization. *Pattern Recognition*, 155:110706, 2024.

Zhe Cao, Tao Zhang, Yichen Wang, Haonan Xin, Rong Wang, and Feiping Nie. Scalable graph-based clustering for unaligned multi-view data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–16, 2026. doi: 10.1109/TPAMI.2026.3729883.

Yuzhuo Dai, Siwei Wang, Zhibin Dong, Xinhang Wan, Tianrui Liu, Kunlun He, Xinwang Liu, and En Zhu. Transport barycenter-guided sample-to-cluster matching for unaligned multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2026. doi: 10.1109/TPAMI.2026.3720198.

Shanghui Deng, Xiao Zheng, Chang Tang, Xinwang Liu, Yuanyuan Liu, and Shan An. scspaf: Cell similarity purified adaptive fusion network for single-cell multi-omics clustering. *IEEE Transactions on Computational Biology and Bioinformatics*, 22(6):2943–2954, 2025.

Shanghui Deng, Xiao Zheng, Chang Tang, Lianbo Guo, and Xinwang Liu. Game theory inspired cross-view interaction alignment for partially view-aligned clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–15, 2026. doi: 10.1109/TPAMI.2026.3702325.

Wenhua Dong, Xiao-Jun Wu, Zhenhua Feng, Sara Atito, Muhammad Awais, and Josef Kittler. Probabilistically aligned view-unaligned clustering with adaptive template selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1904–1916, 2026a.

Zhibin Dong, Shengju Yu, Siwei Wang, Xinwang Liu, and En Zhu. Cross-view graph matching for unsupervised learning with unaligned multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 38(4):2221–2233, 2026b.

Lunke Fei, Junlin He, Qi Zhu, Shuping Zhao, Jie Wen, and Yong Xu. Deep multi-view contrastive clustering via graph structure awareness. *IEEE Transactions on Image Processing*, 34:3805–3816, 2025.

L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pp. 524–531 vol. 2, 2005. doi: 10.1109/CVPR.2005.16.

Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.

Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. Robust contrastive multi-view clustering against dual noisy correspondence. *Advances in Neural Information Processing Systems*, 37:121401–121421, 2024.

Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. Robust variational contrastive learning for partially view-unaligned clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4167–4176, 2024.

Zhenyu Huang, Joey Tianyi Zhou, Xi Peng, Changqing Zhang, Hongyuan Zhu, and Jiancheng Lv. Multi-view spectral clustering network. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 2563–2569. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/356. URL https://doi.org/10.24963/ijcai.2019/356.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Haobin Li, Yijie Lin, Peng Hu, Mouxing Yang, and Xi Peng. Learning with dual-level noisy correspondence for multi-modal entity alignment. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 151700–151731, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f62114abe24195d7d153438b1c0eaf72-Paper-Conference.pdf.

Mengran Li, Ronghui Zhang, Yong Zhang, Xinglin Piao, Shiyu Zhao, and Baocai Yin. Scae: Structural contrastive auto-encoder for incomplete multi-view representation learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9), September 2024a. ISSN 1551-6857. doi: 10.1145/3672078. URL https://doi.org/10.1145/3672078.

Xingfeng Li, Yuangang Pan, Yinghui Sun, Quansen Sun, Ivor Tsang, and Zhenwen Ren. Fast unpaired multi-view clustering. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 4488–4496. International Joint Conferences on Artificial Intelligence Organization, 8 2024b. doi: 10.24963/ijcai.2024/496. URL https://doi.org/10.24963/ijcai.2024/496.

Xingfeng Li, Jiawei Peng, Zhongwen Wang, Yuan Sun, Yuying Zhu, Fei Zhou, and Zhenwen Ren. Energy-preserving shifted bipartite graph learning for unpaired large-scale multi-view clustering. *Neural Networks*, 201:108952, 2026b. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2026.108952. URL https://www.sciencedirect.com/science/article/pii/S0893608026004132.

Ke Liang, Lingyuan Meng, Hao Li, Jun Wang, Long Lan, Miaomiao Li, Xinwang Liu, and Huaimin Wang. From concrete to abstract: Multi-view clustering on relational knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):9043–9060, 2025.

Jiyuan Liu, Xinwang Liu, Siqi Wang, Xinhang Wan, Dongsheng Li, Kai Lu, and Kunlun He. Communication-efficient federated multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):17–32, 2026.

Shen Ouyang, Honglin Yuan, Dongliang Wang, Yuan Sun, Zhenwen Ren, and Xingfeng Li. Clip-driven lifelong multi-view clustering. *Pattern Recognition*, 178:113425, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2026.113425. URL https://www.sciencedirect.com/science/article/pii/S0031320326003900.

Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. COMIC: Multi-view clustering without parameter selection. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5092–5101. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/peng19a.html.

Usha Ruby, Vamsidhar Yendapalli, et al. Binary cross entropy with deep learning technique for image classification. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(10), 2020.

Bohang Sun, Yuena Lin, Tao Yang, Zhen Zhu, Zhen Yang, and Gengyu Lyu. AF-UMC: An alignment-free fusion framework for unaligned multi-view clustering. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=G1jrjumK1b.

Yuan Sun, Yongxiang Li, Zhenwen Ren, Guiduo Duan, Dezhong Peng, and Peng Hu. Roll: Robust noisy pseudo-label learning for multi-view clustering with noisy correspondence. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30732–30741, 2025. doi: 10.1109/CVPR52734.2025.02862.

Huibing Wang, Mingze Yao, Guangqi Jiang, Zetian Mi, and Xianping Fu. Graph-collaborated auto-encoder hashing for multiview binary clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):10121–10133, 2024. doi: 10.1109/TNNLS.2023.3239033.

Yi Wen, Siwei Wang, Qing Liao, Weixuan Liang, Ke Liang, Xinhang Wan, and Xinwang Liu. Unpaired multi-view graph clustering with cross-view structure matching. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16049–16063, 2024.

Baili Xiao, Zhibin Dong, Ke Liang, Suyuan Liu, Siwei Wang, Tianrui Liu, Xingchen Hu, En Zhu, and Xinwang Liu. Easemvc: Efficient dual selection mechanism for deep multi-view clustering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20716–20726, 2025. doi: 10.1109/CVPR52734.2025.01929.

Shunxin Xiao, Shide Du, Zhaoliang Chen, Yunhe Zhang, and Shiping Wang. Dual fusion-propagation graph neural network for multi-view clustering. *IEEE Transactions on Multimedia*, 25:9203–9215, 2023. doi: 10.1109/TMM.2023.3248173.

Like Xin, Wanqi Yang, Lei Wang, and Ming Yang. Multilevel reliable guidance for unpaired multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):18968–18982, 2025.

Like Xin, Wanqi Yang, Lei Wang, and Ming Yang. Global-and-local guidance with synthesized view for unpaired multi-view clustering. *Information Processing & Management*, 63(3):104502, 2026. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2025.104502. URL https://www.sciencedirect.com/science/article/pii/S0306457325004431.

Gehui Xu, Chengliang Liu, Yabo Liu, Yong Xu, and Jie Wen. Partial contrastive learning for partially view-aligned multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):12068–12081, 2026. doi: 10.1109/TPAMI.2026.3694999.

Xueming Yan, Ziqi Wang, and Yaochu Jin. Heterogeneity-aware federated graph neural networks for incomplete multi-view clustering. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 10(1):369–381, 2026.

Chao Zhang, Zhi Wang, Xiuyi Jia, Zechao Li, Chunlin Chen, and Huaxiong Li. Multi-view clustering with incremental instances and views. *IEEE Transactions on Image Processing*, 34:4203–4214, 2025a.

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*, volume 39, pp. 22650–22658, 2025b.

Yuanyang Zhang, Xinhang Wan, Chao Zhang, Jie Xu, Cunjian Chen, Tien-Tsin Wong, Li Yao, and Yijie Lin. Mamba-driven multi-view discriminative clustering via global-local cross-view sequence modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28519–28527, 2026.

Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1774–1782, 2019.

Liang Zhao, Tianqi Yue, Shubin Ma, Ziyue Wang, ZhiYuan Liu, and Bo Xu. Knnda: A new perspective of alignment recovery for partially view-aligned clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28706–28714, 2026a.

Liang Zhao, Tianqi Yue, Shubin Ma, Ziyue Wang, and Lin Yuanbo Wu. Pseudo-label guided alignment propagation for partially view-aligned clustering. *IEEE Transactions on Multimedia*, pp. 1–12, 2026b. doi: 10.1109/TMM.2026.3727807.

Haochen Zhou, Guofeng Ding, Mouxing Yang, Peng Hu, Yijie Lin, and Xi Peng. Uncover underlying correspondence for robust multi-view clustering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=a4S1nQay3b.

# APPENDIX

## A OPTIMIZATION PROCEDURE AND DATASET DETAILS

### A.1 END-TO-END OPTIMIZATION PROCEDURE

Algorithm 1 summarizes the complete training and inference pipeline, including masked topology reconstruction, progressive pseudo-correspondence mining, cross-view feature synchronization, instance-to-cluster alignment, and final soft fusion.

**Algorithm 1** Pseudo-Correspondence Guided Blind-Alignment Learning

**Require:** Multi-view data $\{\mathbf{X}^v\}_{v=1}^V$, cluster number $K$, template view $t$, number of epochs $C$ tunable parameters $\lambda_1, \lambda_2, p, \kappa$
**Ensure:** Clustering assignments $\mathbf{y}$

```
Construct k-NN graphs and initialize the model
for c = 0 to C − 1 do
    Obtain view-specific representations {Z^v} via Eq. (1)
    Compute the reconstruction loss L_rec via Eq. (2)
    Compute H^v = ℓ2-norm(P(Z^v)) for all v
    for v ≠ t do
        Mine self-paced pseudo-correspondences A^(c) via Eqs. (3)–(4)
        Compute L_fs^(t,v) and L_i2c^(t,v) via Eqs. (5) and (9)
    end for
    Compute L_total = L_rec + λ1 L_fs + λ2 L_i2c via Eq. (11)
    Update model parameters by backpropagation
end for
Softly align auxiliary-view representations to the template view via Eq. (12)
Fuse Z_fuse = Concat(H^t, {Ĥ^v}_{v≠t})
Apply k-means to Z_fuse to obtain clustering assignments y
```

### A.2 DATASET STATISTICS

Table 4 summarizes the sample size, number of views, number of clusters, and feature dimensionality of each benchmark used in the experiments.

Table 4: Statistics of the eight benchmark datasets.

| Dataset | Samples | Views | Clusters | Dimensions |
|---|---|---|---|---|
| Handwritten | 2000 | 6 | 10 | 216/76/64/6/240/47 |
| Scene15 | 4485 | 3 | 15 | 1800/1180/1240 |
| MNIST-USPS | 5000 | 2 | 10 | 784/256 |
| GSE | 8200 | 2 | 27 | 25/1000 |
| Hdigit | 10000 | 2 | 10 | 784/256 |
| ALOI | 10800 | 4 | 100 | 77/13/64/125 |
| ReutersEF | 18758 | 2 | 6 | 100/100 |
| Cifar10 | 50000 | 3 | 10 | 512/2048/1024 |

Table 5: Additional ablation results on MNIST-USPS, GSE, Hdigit, ALOI, ReutersEF, and Cifar10. The best results are shown in bold. 'Pur', 'Fsc', and 'Pre' denote Purity, F-score, and Precision, respectively.

| Variant | MNIST-USPS | | | | | | GSE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | NMI | Pur | Fsc | Pre | ARI | ACC | NMI | Pur | Fsc | Pre | ARI |
| w/o MTMM | 64.07 | 62.10 | 67.02 | 63.62 | 66.06 | 50.20 | 66.46 | 75.03 | 71.37 | 61.44 | 64.30 | 59.46 |
| w/o SPS | 68.98 | 65.27 | 70.77 | 68.85 | 72.18 | 53.66 | 66.64 | 75.25 | 70.33 | 62.28 | 64.62 | 60.17 |
| w/o MP | 65.26 | 63.09 | 69.86 | 64.96 | 68.31 | 50.75 | 67.05 | 76.17 | 72.11 | 63.39 | 66.62 | 60.59 |
| w/o CFS | 69.60 | 65.71 | 71.43 | 69.31 | 72.00 | 54.66 | 68.18 | 77.08 | 71.83 | 63.50 | 66.42 | 61.38 |
| w/o I2C-CA | 63.25 | 62.83 | 67.94 | 63.07 | 67.01 | 49.37 | 68.49 | 77.20 | 71.23 | 63.41 | 65.76 | 62.59 |
| **Full** | **73.48** | **67.93** | **73.48** | **73.33** | **76.70** | **57.74** | **73.40** | **77.95** | **76.22** | **68.37** | **70.74** | **64.68** |
| **Variant** | **Hdigit** | | | | | | **ALOI** | | | | | |
| | ACC | NMI | Pur | Fsc | Pre | ARI | ACC | NMI | Pur | Fsc | Pre | ARI |
| w/o MTMM | 66.73 | 62.42 | 68.45 | 66.23 | 69.16 | 50.25 | 51.06 | 69.07 | 53.64 | 49.82 | 51.57 | 38.56 |
| w/o SPS | 66.84 | 65.22 | 69.60 | 66.70 | 69.67 | 52.84 | 51.20 | 69.04 | 53.49 | 49.83 | 51.70 | 38.77 |
| w/o MP | 62.62 | 63.05 | 66.94 | 62.98 | 66.01 | 49.91 | 47.97 | 64.09 | 49.81 | 47.51 | 50.10 | 32.91 |
| w/o CFS | 67.67 | 64.81 | 70.56 | 67.28 | 70.08 | 52.92 | 50.86 | 69.17 | 53.25 | 49.50 | 51.44 | 38.54 |
| w/o I2C-CA | 65.97 | 64.77 | 69.60 | 65.77 | 68.86 | 52.32 | 48.65 | 69.67 | 51.48 | 46.15 | 47.94 | 38.68 |
| **Full** | **74.61** | **67.76** | **74.61** | **74.28** | **75.98** | **59.48** | **54.13** | **69.72** | **55.73** | **52.90** | **53.99** | **40.93** |
| **Variant** | **ReutersEF** | | | | | | **Cifar10** | | | | | |
| | ACC | NMI | Pur | Fsc | Pre | ARI | ACC | NMI | Pur | Fsc | Pre | ARI |
| w/o MTMM | 53.38 | 33.96 | 59.82 | 48.36 | 51.55 | 28.06 | 80.92 | 75.04 | 80.92 | 80.15 | 80.63 | 70.94 |
| w/o SPS | 54.41 | 35.88 | 60.89 | 48.69 | 51.12 | 28.96 | 80.35 | 74.50 | 80.38 | 80.31 | 81.79 | 69.72 |
| w/o MP | 56.83 | 36.45 | 61.41 | 52.03 | 53.76 | 28.82 | 76.97 | 75.78 | 79.26 | 76.97 | 79.46 | 69.40 |
| w/o CFS | 54.15 | 35.42 | 60.65 | 47.56 | 49.48 | 28.45 | 72.69 | 65.80 | 73.50 | 71.46 | 72.41 | 58.05 |
| w/o I2C-CA | 53.79 | 33.55 | 59.82 | 50.98 | 54.28 | 27.36 | 80.57 | 76.20 | 80.57 | 81.12 | 82.73 | 71.20 |
| **Full** | **65.87** | **43.54** | **68.21** | **61.64** | **63.04** | **37.96** | **89.13** | **78.63** | **89.13** | **89.33** | **90.02** | **77.33** |

### A.3 Computational Complexity Analysis

Let $N$, $V$, $d_v$, $h$, $d$, $K$, $k$, $\kappa$, and $C$ denote the number of instances, number of views, input dimension, hidden dimension, latent dimension, number of clusters, number of graph neighbors, number of candidate neighbors, and number of training epochs, respectively. Constructing the $k$-NN graphs costs $O(N^2 \sum_{v=1}^{V} d_v)$. In each epoch, representation learning requires $O(Nh \sum_{v=1}^{V} d_v + VNhd + VNk(h+d))$, pseudo-correspondence mining requires $O((V-1)N^2 d + (V-1)N\kappa\sqrt{N})$, feature synchronization requires $O((V-1)N_a^2 d)$, and instance-to-cluster alignment requires $O((V-1)N_a K d)$, where $N_a$ is the number of selected pseudo-correspondences. Since $N_a = O(N)$ and $K$, $k$, and $\kappa$ are small constants, the dominant dependence on $N$ is approximately $O(CVN^2 d)$. During inference, soft alignment costs $O((V-1)N^2 d)$, while $k$-means requires $O(TNKVd)$ for $T$ iterations. The space complexity is $O(N \sum_{v=1}^{V} d_v + VNk + VNd + N^2)$ when auxiliary views are processed sequentially, which is dominated by $O(N^2)$; storing all cross-view matrices simultaneously increases it to $O(VN^2)$.

## B Additional Ablation Results (**Q2**)

Table 5 reports the additional ablation results on MNIST-USPS, GSE, Hdigit, ALOI, ReutersEF, and Cifar10. The results exhibit trends consistent with those reported in the main text, further confirming the effectiveness of each component across different datasets.

# C STATISTICAL VALIDATION OF THE MAIN RESULTS (Q4)

## C.1 CROSS-DATASET RADAR-CHART ANALYSIS

Figure 3 provides a unified comparison of clustering performance across the eight datasets. Each radar chart corresponds to one evaluation metric, with all datasets arranged consistently across the six panels. Our method forms the outer performance envelope on nearly all axes, achieving the best result in 46 of the 48 dataset–metric combinations. The only exceptions are NMI and Purity on ALOI, where AFUMC performs better. Overall, the consistently broad radar profiles demonstrate the strong and stable performance of our method across different datasets and evaluation metrics.

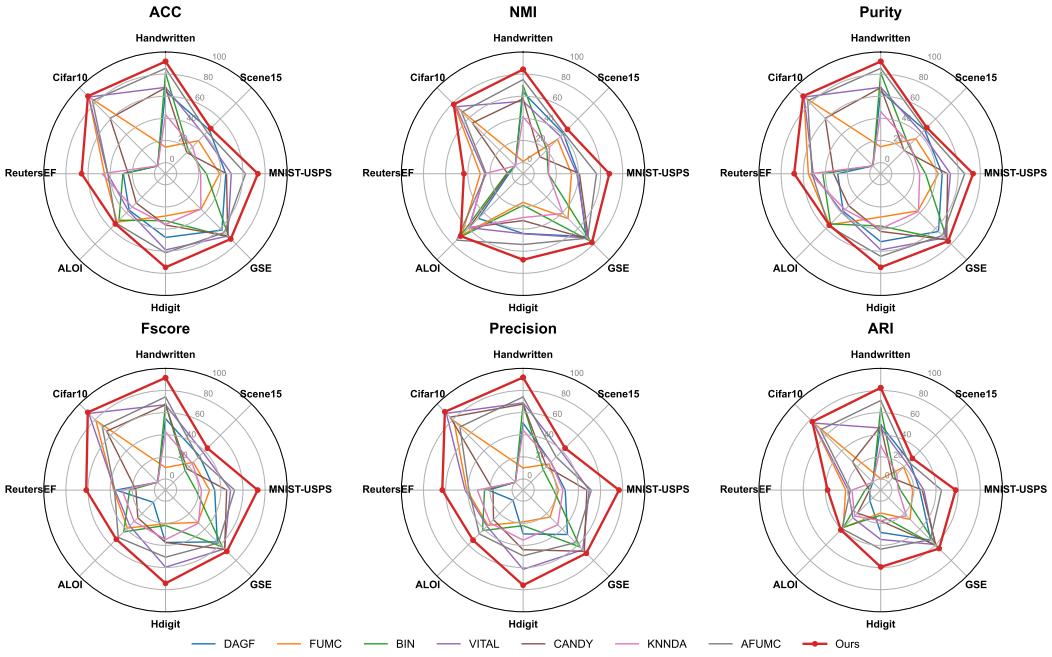


Figure 3: Radar-chart comparison of eight methods across eight datasets in terms of ACC, NMI, Purity, F-score, Precision, and ARI.

## C.2 PAIRWISE STATISTICAL SIGNIFICANCE

To further evaluate the statistical reliability of the improvements, we conduct one-sided Wilcoxon signed-rank tests between our method and each baseline on all six metrics, using datasets with finite results for both methods as paired observations. For DAGF, BIN, and KNNDA, Cifar10 is excluded because these baselines yield NaN results, giving an effective sample size of $n = 7$; all other comparisons use all eight datasets ($n = 8$). As shown in Table 6, our method significantly outperforms all competing methods at the $p < 0.05$ level. Comparisons with $n = 8$ mostly reach $p < 0.005$, while the relatively weaker statistical significance against AFUMC on NMI and Purity is consistent with its strong performance on ALOI.

# D T-SNE REPRESENTATION VISUALIZATION (Q5)

To complement the quantitative clustering metrics, we visualize the final representations learned under the completely unaligned Handwritten setting using t-SNE. As shown in Figure 4, FUMC and VITAL produce rather disordered cluster distributions. In contrast, our representations exhibit more coherent class-wise organization, consistent with the quantitative improvements reported in the main experiments.

Table 6: One-sided Wilcoxon signed-rank tests of our method against the competing methods. Significance levels are denoted by $\circ$ for $p < 0.05$, $\triangle$ for $p < 0.01$, and $\star$ for $p < 0.005$. For each baseline, only datasets with finite results for both methods are retained as paired observations; thus, Cifar10 is excluded for DAGF, BIN, and KNNDA. The effective number of paired observations is reported as $n$.

| Ours vs. Method | $n$ | ACC | NMI | Purity | F-score | Precision | ARI |
|---|---|---|---|---|---|---|---|
| DAGF | 7 | $\triangle$ | $\triangle$ | $\triangle$ | $\triangle$ | $\triangle$ | $\triangle$ |
| FUMC | 8 | $\star$ | $\star$ | $\star$ | $\star$ | $\star$ | $\star$ |
| BIN | 7 | $\triangle$ | $\circ$ | $\triangle$ | $\triangle$ | $\triangle$ | $\triangle$ |
| VITAL | 8 | $\star$ | $\star$ | $\star$ | $\star$ | $\star$ | $\star$ |
| CANDY | 8 | $\star$ | $\star$ | $\star$ | $\star$ | $\star$ | $\star$ |
| KNNDA | 7 | $\triangle$ | $\triangle$ | $\triangle$ | $\triangle$ | $\triangle$ | $\triangle$ |
| AFUMC | 8 | $\star$ | $\circ$ | $\triangle$ | $\star$ | $\star$ | $\star$ |

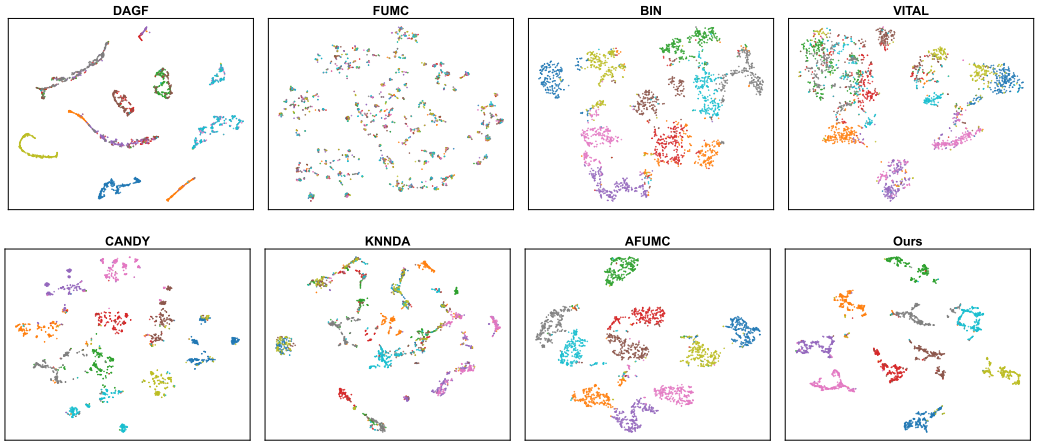


Figure 4: t-SNE visualization of the final Handwritten representations under the completely unaligned setting.

## E Template-View Selection and Hyperparameter Analysis (**Q6**)

### E.1 Template-View Selection Strategy

We investigate how the choice of template view affects clustering performance by exhaustively evaluating every available view as the template on Handwritten, Scene15, MNIST-USPS, and Hdigit. We additionally identify the view selected by a data-driven strategy that independently performs K-means on the complete samples of each view and chooses the view with the maximum silhouette coefficient. In Table 7, the corresponding candidate is labeled as Max-Silhouette.

The first-view strategy outperforms maximum-silhouette selection across all four datasets and all six reported metrics. On average, choosing the first view instead of the maximum-silhouette view improves ACC and ARI by 12.44 and 13.87 percentage points, respectively. However, the exhaustive comparison also shows that the first view is not universally optimal: on Handwritten, View 5 achieves the best result on every metric, improving ACC from 91.45% to 94.40% and ARI from 82.47% to 88.10% relative to View 1. In contrast, View 1 remains the best candidate on Scene15, MNIST-USPS, and Hdigit. These results demonstrate that template quality is dataset-dependent and that within-view clustering quality alone is insufficient for template selection, as the silhouette coefficient does not capture cross-view semantic compatibility, correspondence recoverability, or pseudo-pair reliability.Accordingly, the first view is retained as a strong and consistent default in the remaining experiments, while automatic template selection remains a promising direction for future work.

Table 7: Comparison of individual template views and the maximum-silhouette selection strategy. The best result for each dataset and metric is shown in bold. Pur, Fsc, and Pre denote Purity, F-score, and Precision, respectively.

| Dataset | Template Strategy | ACC | NMI | Pur | Fsc | Pre | ARI |
|---|---|---|---|---|---|---|---|
| Handwritten | View 1 | 91.45 | 84.26 | 91.45 | 91.48 | 91.80 | 82.47 |
| | View 2 | 71.25 | 65.44 | 71.25 | 71.19 | 73.16 | 54.61 |
| | View 3 | 87.95 | 81.08 | 87.95 | 87.95 | 88.08 | 76.87 |
| | Max-Silhouette (View 4) | 69.75 | 67.32 | 69.75 | 68.95 | 70.25 | 55.62 |
| | View 5 | **94.40** | **89.20** | **94.40** | **94.41** | **94.63** | **88.10** |
| | View 6 | 71.90 | 67.60 | 71.90 | 71.01 | 71.44 | 58.48 |
| Scene15 | View 1 | **47.76** | **46.67** | **48.92** | **43.57** | **43.63** | **30.54** |
| | Max-Silhouette (View 2) | 37.55 | 37.92 | 38.44 | 37.02 | 38.47 | 21.09 |
| | View 3 | 34.09 | 33.77 | 37.12 | 33.68 | 38.46 | 17.26 |
| MNIST-USPS | View 1 | **73.48** | **67.93** | **73.48** | **73.33** | **76.70** | **57.74** |
| | Max-Silhouette (View 2) | 65.46 | 61.05 | 65.46 | 65.27 | 65.56 | 48.67 |
| Hdigit | View 1 | **74.61** | **67.76** | **74.61** | **74.28** | **75.98** | **59.48** |
| | Max-Silhouette (View 2) | 64.79 | 62.21 | 65.59 | 64.82 | 65.36 | 49.39 |

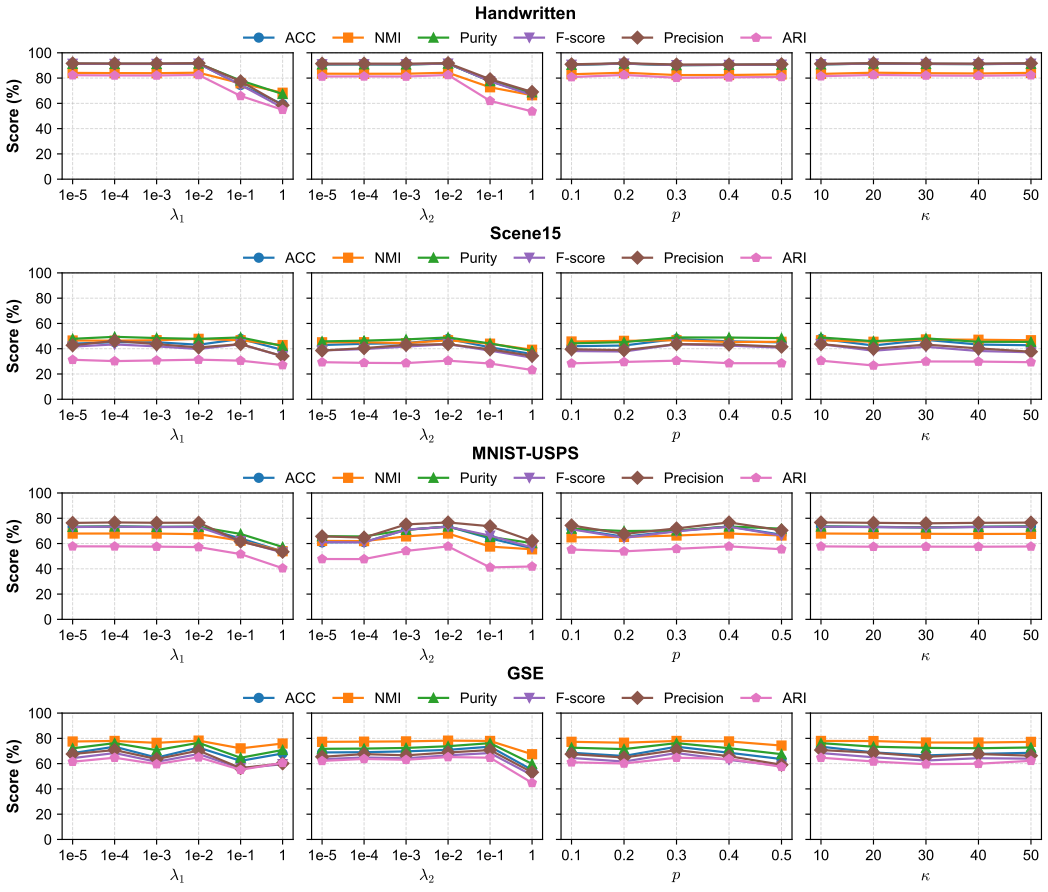


Figure 5: Parameter sensitivity analysis of $\lambda_1$, $\lambda_2$, $\kappa$, and $p$ on Handwritten, Scene15, MNIST-USPS, and GSE.

### E.2 Parameter Sensitivity Analysis

We analyze the sensitivity of four principal hyperparameters on Handwritten, Scene15, MNIST-USPS, GSE, Hdigit, ALOI, ReutersEF, and Cifar10. In each experiment, only one hyperparameter

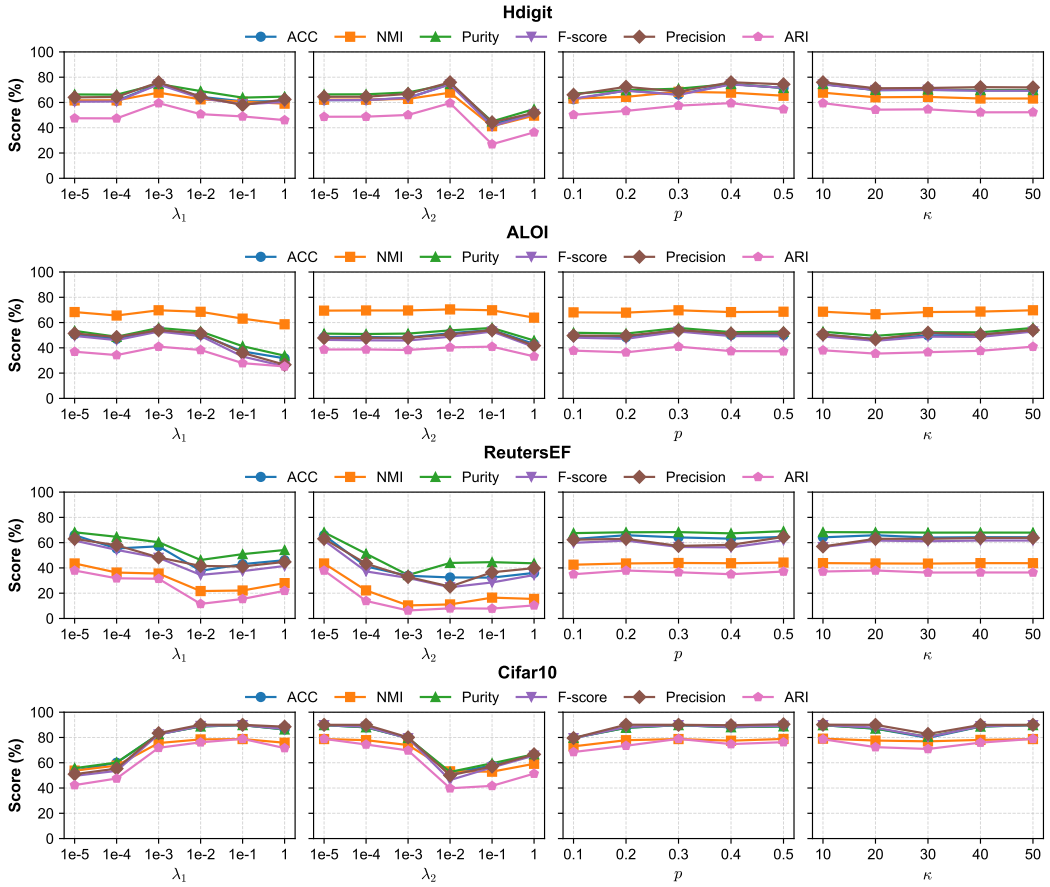


Figure 6: Parameter sensitivity analysis of $\lambda_1$, $\lambda_2$, $\kappa$, and $p$ on Hdigit, ALOI, ReutersEF, and Cifar10.

is varied while the remaining settings are fixed to the dataset-specific configuration obtained from the full grid search. Specifically, we evaluate $\lambda_1, \lambda_2 \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, $\kappa \in \{10, 20, 30, 50\}$, and $p \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. The complete curves of ACC, NMI, Purity, F-score, Precision, and ARI are shown in Figures 5 and 6.

Overall, $\lambda_1$ and $\lambda_2$ exhibit certain dataset dependence, while excessively large loss weights generally lead to degraded performance. Therefore, $\lambda_1$ and $\lambda_2$ are recommended to be searched over $[10^{-2}, 10^{-1}]$, which provides a suitable balance among feature synchronization, cluster-level alignment, and representation learning.

In comparison, the candidate-neighborhood size $\kappa$ is relatively stable within the tested range, and $\kappa \in [10, 30]$ is recommended for most datasets. For the topology masking ratio $p$, moderate masking strength generally yields more stable performance, with $p = 0.3$ serving as a reasonable default choice. Overall, these observations provide practical guidance for hyperparameter configuration and help narrow the search space when applying the proposed framework to new datasets.

# F Optimization Behavior and Computational Efficiency

## F.1 Convergence Analysis (**Q7**)

The optimization behavior of the proposed framework under the completely unaligned setting is further investigated. Figure 7 reports the average total objective $\mathcal{L}_{total}$ at each training epoch. Across all eight datasets, the objective decreases rapidly during the early training stage and then gradu-

ally approaches a stable plateau. This consistent convergence behavior indicates that the jointly optimized components can be trained stably even without any known cross-view correspondences.

### F.2 Running Time Comparison (**Q8**)

To evaluate the practical running cost of the proposed framework, the running times of all methods are compared under the completely unaligned setting. Figure 8 reports $\log_{10}(\text{Time} + 1)$, where Time denotes the running time in seconds. Missing results on Cifar10 are displayed as zero solely for visualization. Across most datasets, the proposed method exhibits a moderate running time and is generally faster than AFUMC, KNNDA, and CANDY. The proposed method does not achieve the lowest running time. However, it maintains a favorable balance between computational cost and clustering performance.

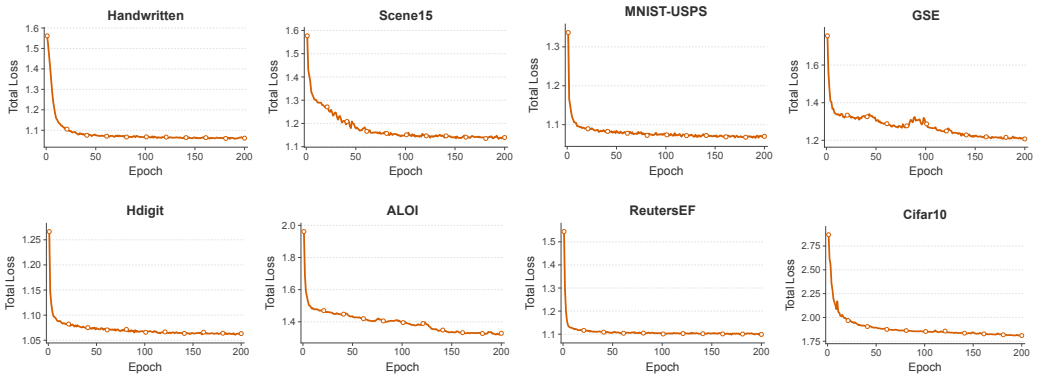


Figure 7: Convergence curves of the total training objective on all datasets under the completely unaligned setting.

## G Robustness to Noise (**Q9**)

To evaluate robustness to feature corruption, we conduct Gaussian noise experiments on Handwritten under the completely unaligned setting. Specifically, an increasing proportion of samples in each view is independently corrupted with Gaussian outlier noise, introducing progressively stronger perturbations to the original feature space. Figure 9 reports six clustering metrics under different noise ratios and additionally includes the w/o MTMM variant to examine the role of masked topological modeling under noisy conditions.

As the noise ratio increases, the performance of all methods generally decreases due to the growing corruption of local feature structures. Nevertheless, the full model consistently maintains better clustering performance than the competing methods across different noise levels, indicating its ability to preserve reliable cross-view information under feature perturbations. Moreover, the performance advantage becomes more evident at higher noise ratios, suggesting that PCGBAL is less sensitive to severe feature corruption.

The comparison with w/o MTMM further highlights the importance of masked topological reconstruction. Without topology masking, the model directly learns from the complete $k$-NN graph and is therefore more susceptible to unreliable neighborhood relations introduced by noisy samples. In contrast, MTMM randomly masks graph connections and reconstructs them from the remaining visible topology, discouraging excessive dependence on individual local relations and promoting more stable structural representations. Consequently, the full model exhibits substantially better robustness as the noise level increases.

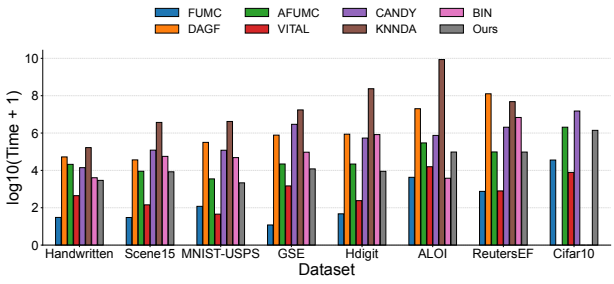


Figure 8: Running-time comparison under the completely unaligned setting.

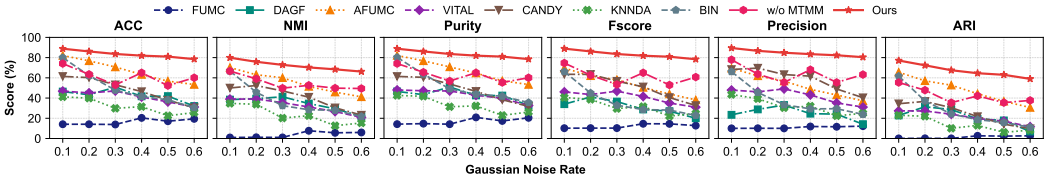


Figure 9: Clustering performance on Handwritten under increasing Gaussian outlier noise in the completely unaligned setting.