MITIGATING MODALITY CONFLICTS IN MULTI-MODAL CLUSTERING WITH DECOUPLED NEIGHBORHOODS

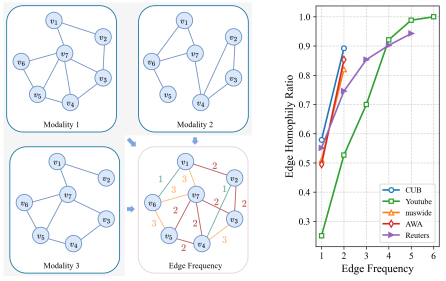
Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal clustering aims to discover underlying cluster structures by exploring information from different modalities. However, existing methods often struggle with conflicts across diverse modalities, where modality-specific information often leads to inconsistent topological relationships. To address this issue, we propose a novel neighborhood decoupling strategy that decomposes multi-modal graphs into a set of edge-disjoint graphs based on cross-modal edge frequencies. This is driven by the insight that edges repeatedly supported by multiple modalities exhibit higher semantic reliability. By stratifying the structural information into a hierarchy of decoupled neighborhoods according to edge frequencies, we introduce a structure alignment objective that enforces a structural ordering and injects the hierarchical topology into the modality-joint representation for clustering. Furthermore, we establish theoretical guarantees showing that optimizing this objective minimizes an upper bound on the k-means cost, ensuring compact cluster structures. Extensive experiments are conducted to verify the superiority of the proposed method compared with state-of-the-art approaches on multiple benchmark datasets.

1 INTRODUCTION

With the development of sensor technologies, the growing number of multi-modal data has driven the demand for machine learning algorithms capable of reasoning over diverse information sources, such as text, audio, and visual data Li et al. (2025). In the unsupervised regime, Multi-modal clustering (MMC) plays a key role by partitioning data points into distinct semantic groups by leveraging the information embedded across modalities, and has a wide range of applications in computer vision and machine learning Li et al. (2026b). However, effectively fusing this information remains a fundamental challenge due to inherent *modality conflicts* arising from modality-unique information.



(a) (b)

Figure 1: (a) Illustration of edge frequencies in multi-modal data. (b) Edge homophily ratios of neighborhoods under different edge frequencies.

Most existing approaches alleviate modality conflicts through information disentanglement, decomposing multi-modal information into modality-consistent (e.g., information across different modalities aligning with the underlying cluster structures) and modality-specific components Huang et al. (2022; 2021); Nie et al. (2021); Huang et al. (2023b); Zou et al. (2026); Cui et al. (2023). However, this binary decomposition can be overly restrictive because the boundary between the two is intrinsically blurred. Enforcing a hard separation may either discard subtle but discriminative cross-modal information or retain modality-specific noise, thereby limiting clustering performance.

In this paper, we shift the paradigm from binary separation to fine-grained neighborhood decoupling. As illustrated in Fig. 1 (a), three graphs corresponding to three modalities contain different edges. Our key insight is that the frequency of an edge appearing across multiple modality graphs serves as a natural and robust proxy for its reliability. This intuition is empirically grounded in Fig. 1(b),

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.

where our statistical analysis reveals a strong positive correlation between edge frequency and the homophily ratio.

Motivated by this insight, we propose a neighborhood decoupling strategy that reformulates the multi-modal graphs into multiple edge-disjoint ones based on edge frequencies across different modalities. Compared with the binary consistent vs. specific decomposition paradigm, our strategy organizes topological information into a hierarchy of neighborhoods with varying degrees of consistency with the underlying cluster structures. In this case, it enables the model to adaptively leverage high-frequency edges while gracefully incorporating lower-frequency edges as auxiliary information, offering a more flexible and robust way to tackle modality conflicts in multi-modal clustering. Building upon this perspective, we propose Decoupled Neighborhood-Induced Multi-Modal Clustering (DNI-MMC). It first employs the aforementioned decoupling strategy to decompose multi-modal graphs into hierarchical levels neighborhoods. To effectively leverage the prior information of decoupled neighborhoods, we design a novel structure alignment loss and its tractable surrogate form to inject the hierarchical neighborhood information into the modality-joint representation for clustering. We highlight our contributions as follows:

- We propose DNI-MMC, a decoupled neighborhood-induced multi-modal clustering algorithm that learns a structural representation for clustering multi-modal data with modality conflicts.
- We introduce a neighborhood decoupling strategy to decompose the multi-modal graphs into edge-disjoint ones by stratifying the edge frequencies across different modalities to mitigate modality conflicts.
- We design a structure alignment loss and derive its tractable surrogate form to preserve the decoupled neighborhood topology within the modality-joint representation. We provide a theoretical analysis demonstrating that this optimization explicitly encourages compact cluster structures.
- We validate DNI-MMC across multiple benchmarks, showing state-of-the-art performance on both multi-modal clustering and multi-modal attribute graph clustering tasks.

2 Related Work

There are a number of works on effective decoupling of multi-modal information for MMC. The main pipeline of these methods is to separate the multi-modal information into modality-specific and modality-consistent parts based on different criteria, such as connected components in a graph Huang et al. (2022; 2021); Nie et al. (2021); Tang et al. (2025) or information-theoretic measures Huang et al. (2023b); Cui et al. (2023). For instance, Huang et al. Huang et al. (2022) decomposed multiple subspace representations into consistent and diverse parts, and searched for a structural graph from the consistent part under an exact connected components constraint. In Huang et al. (2023b), an MMC method is proposed from an information-theoretic perspective, where three requirements (i.e., comprehensiveness, concentration, and cross-diversity) are introduced to achieve multi-modal learning. However, dividing multi-modal information solely into modality-consistent or modality-specific components is overly restrictive, and this causes the modality-consistent part with insufficient multi-modal information to expose the cluster structures.

Our work also relates to a large and growing number of MMC methods built upon contrastive learning. The core objective of contrastive learning is to separate representations of negative pairs while increasing similarity among positive ones Khosla et al. (2020); Krishnan et al. (2022); Gui et al. (2024). Researchers developed contrastive MMC methods mainly follow the data argumentation-based contrastive learning strategy, which treat the same sample from different views as positive pairs and different samples across diverse views as negative pairs Yang et al. (2023); Chen et al. (2023); Trosten et al. (2023). The primary purpose of these methods is to align multi-modal information in the feature space, facilitating the fusion procedure of multi-modal information for clustering Xu et al. (2022); Lin et al. (2021); Luo et al. (2024); Chen et al. (2023); Xi et al. (2026). However, they lack of explicitly utilizing the neighboring information among samples, and are insufficient to guarantee obtaining a structural representation for clustering. Another line of works related to ours are the graph contrastive methods without augmentations, which simply consider the neighboring nodes as positive pairs Xiao et al. (2023; 2024); Zhang et al. (2022). However, these models are designed for a single graph and are difficult to directly extend to the MMC task.

Finally, there are studies of MMC based on collaborative training strategies Xu et al. (2021); Wang et al. (2024), co-regularized learning Wang et al. (2018; 2020); Kang et al. (2020); Nie et al. (2018); Wang et al. (2023a); Zhang et al. (2024), and adaptive multi-view fusion schemes Wu et al. (2024); Zhu et al. (2024); Zou et al. (2026). Notably, these methods either rarely address modality inconsistency or rely on a binary consistency-specificity decomposition paradigm to decouple multimodal information for mitigating modality conflicts, highlighting the fundamental differences in the design principles of models.

3 PRELIMINARIES

Problem Setup. Throughout the paper, the multi-modal data is denoted as $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(\ell)}$, where ℓ is the number of modalities. Matrix $\mathbf{X}^{(i)} \in \mathbb{R}^{n \times d_i}$ represents the data of i-th modality, where n and d_i are the number of samples and feature dimensions, respectively. For each modality, we construct a modality-specific nearest neighbor graph $\mathbf{W}^{(i)} \in \{0,1\}^{n \times n}$ with ζ nearest neighbors for each sample, where $\mathbf{W}^{(i)}_{u,v} = 1$ indicates that sample u is connected to sample v in i-th modality.

In this paper, we consider learning a modality-joint representation $\mathbf{F} \in \mathbb{R}^{n \times d}$ from multi-modal data to obtain the clusters $\mathcal{A}_1, \ldots, \mathcal{A}_k$, with the guidance of a union of modality-specific graphs, where d is the number of features, k is the number of clusters. However, different modalities contain their modality-specific information due to different acquisition processes. As a result, an edge that is informative in one modality may be unreliable or even misleading in another. Such *structural inconsistency* hinders the learning of a modality-joint representation for clustering.

Edge Homophily Ratio. To quantify the semantic consistency of a graph with the underlying truth clusters, we adopt the edge homophily ratio, which measures as the proportion of edges connecting two samples that belong to the same cluster. Let $\mathcal{C}_1, \ldots, \mathcal{C}_k$ be the underlying truth clusters and $\mathbf{y} \in \mathbb{R}^n$ be a cluster indicator vector, where $y_u = j$ denotes u-th sample belongs to cluster $\mathcal{C}_j$. Formally, for a graph $\mathbf{W} \in \{0,1\}^{n \times n}$, the edge homophily ratio is defined as

$$h = \mathcal{H}(\mathbf{W}) = \frac{\sum_{u=1} \sum_{v \in \mathcal{N}_u} \mathbb{1}[y_u = y_v]}{\sum_{u=1} |\mathcal{N}_v|} = \frac{\sum_{u=1} \sum_{v \in \mathcal{N}_u} \mathbb{1}[y_u = y_v]}{\mathrm{vol}_{\mathbf{W}}}, \tag{1}$$

where $\mathrm{vol}_{\mathbf{W}}$ is the volume of graph $\mathbf{W}$, $\mathbb{1}[\cdot]$ is an indicator function, $\mathcal{N}_u$ is the neighbor set of u-th sample induced by $\mathbf{W}$. A larger value of $\mathcal{H}(\mathbf{W})$ indicates that the graph is more consistent with the underlying semantic correlations.

4 METHODOLOGY

In this section, we present the details of our multi-modal clustering algorithm. We first give the details of the neighborhood decoupling strategy and discuss why this strategy works for mitigating the modality conflicts. Then, we introduce the optimization goal by connecting the neighborhood-decoupled graphs and the learning process of modality-joint representation. Finally, we establish the structural guarantees for the learned modality-joint representation from the perspective of structure compactness.

4.1 MITIGATING MODALITY CONFLICTS VIA NEIGHBORHOOD DECOUPLING

In multi-modal clustering, distinct modalities frequently exhibit their modality-specific structural preferences, e.g., edge (u, v) is considered a positive connection in modality A, but regarded as noise or dissimilar in modality B. We conceptualize multi-modal graphs as voters that signify the extent of support for an edge across different modalities. This gives a key intuition behind our neighborhood decoupling: edges consistently supported by multiple modalities are more likely to reflect intrinsic semantic relationships.

As shown in Figure 2, given a set of multi-modal graphs $\{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \ldots, \mathbf{W}^{(\ell)}\}$, we stratify the multi-modal graph structures according to the cross-modal occurrence frequency of edges. Formally,

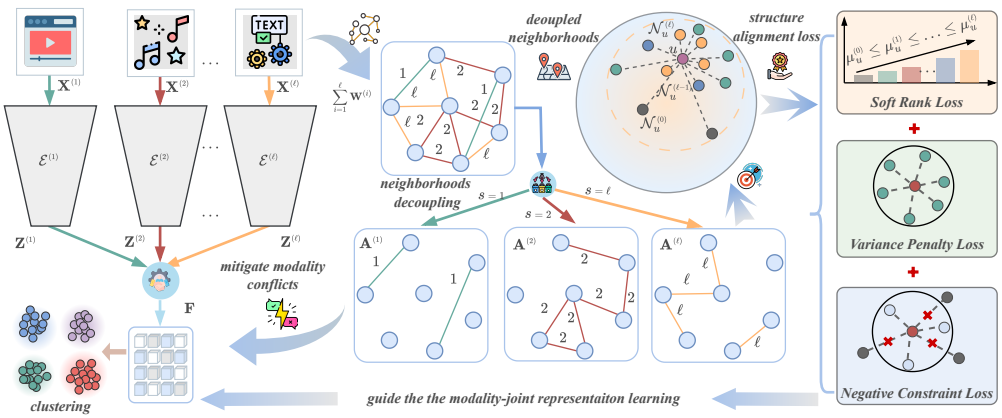


Figure 2: Illustration of decoupled neighborhood-induced multi-modal clustering (DNI-MMC). The modality-joint features are extracted from multi-modal data. Next, the multi-modal graphs are decoupled according to the edge frequencies. Finally, a structural alignment objective is utilized to guide the learning process of modality-joint representation for clustering.

we define the decoupled graph at frequency level s for $s \in [\ell]$ as $\mathbf{A}^{(s)} \in \{0,1\}^{n\times n}$, where

$$\mathbf{A}^{(s)}_{u,v} = \mathbb{1}\left(\sum_{i=1}^{\ell} \mathbf{W}^{(i)}_{u,v} = s\right), \tag{2}$$

That is, $\mathbf{A}^{(s)}_{u,v} = 1$ if and only if the edge (u, v) appears in exactly s modality-specific graphs.

The decomposition partitions the original multi-modal graphs into ℓ disjoint frequency-aware ones $\{\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(\ell)}\}$. Considering the edge frequency as a proxy for structural reliability with respect to the underlying true semantic correlations, the neighborhood decoupling process divides the neighbors of each sample into different subsets corresponding to different reliability levels. By stratifying the neighborhoods according to frequency s, our decoupling strategy alleviates modality conflicts by yielding a hierarchy of neighborhoods. Specifically, high-frequency edges capture the shared consensus across modalities, while modalities that contain conflicting structural information exhibit lower consistency with the underlying true semantic correlations in the resultant decoupled graphs with lower edge frequencies.

4.2 Learning Representation from Decoupled Neighborhoods

In this paper, we use modality-specific feature extractors, i.e., $\mathcal{E}^{(i)}(\mathbf{X}^{(i)}; \Phi^{(i)})$ for $i \in [\ell]$ to capture the modality-specific features from i-th modality as $\mathbf{Z}^{(i)} = \mathcal{E}^{(i)}(\mathbf{X}^{(i)}; \Phi^{(i)})$. After that, the modality-specific features are aggregated into the modality-joint representation $\mathbf{F}$ with $\|\mathbf{f}_u\|_2 = 1$, e.g., $\mathbf{F} = \mathcal{G}\left([\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(\ell)}]; \Theta\right)$. Based on the decoupled graphs, we aim to learn $\mathbf{F}$ from a hierarchy of neighborhoods. Noticing that higher-frequency decoupled graphs are expected to be more reliable, their induced neighbors should receive stronger similarity constraints in the learned representation space. To achieve this, it is intuitive to assign weights to the decoupled sub-graphs based on their structural reliability.

Let $\{\beta_i\}_{i=1}^{\ell}$ be a positive weight sequence with $0 < \beta_1 \leq \beta_2 \leq \ldots \leq \beta_\ell \leq 1$. We introduce a structural alignment objective $\mathcal{L}_{alig}$ for learning $\mathbf{F}$ by weighted aligning the embedding similarities with decoupled neighborhoods, given by

$$\mathcal{L}_{alig} = \sum_{u=1}^{n}\sum_{i=1}^{\ell}\sum_{v\in\mathcal{N}_u^{(i)}} (\beta_i - \mathbf{f}_u^\top \mathbf{f}_v)^2 + \sum_{u=1}^{n}\sum_{v\in\mathcal{N}_u^{(0)}} (\mathbf{f}_u^\top \mathbf{f}_v)^2, \tag{3}$$

where $\mathcal{N}_u^{(i)}$ is the set of nearest neighbors of sample u in $\mathbf{A}^{(i)}$, $\mathcal{N}_u^{(0)}$ denotes the set of samples outside the nearest neighborhoods of the u-th sample across all modalities. In Eq. equation 3, the ordered weights $\beta_1 \leq \beta_2 \leq \ldots \leq \beta_\ell$ force the pairwise similarities of $\mathbf{F}$ to keep the same ordering structures,

along with the neighborhood constraints, such that prior information of decoupled neighborhoods and their relative reliability can be preserved in $\mathbf{F}$.

While $\mathcal{L}_{alig}$ explicitly aligns pairwise similarities with the reliability of different decoupled neighborhoods, it is difficult to optimize directly in unsupervised scenarios. The primary challenge lies in the fact that the optimal weights $\{\beta_i\}_{i=1}^{\ell}$ are *unobservable*, as they are intrinsically coupled with the edge homophily ratios $\{h_i\}_{i=1}^{\ell}$ for ℓ decoupled graphs. To this end, we shift the optimization focus from precise coordinate alignment of pairwise similarities to *ordinal relationship preservation*. Instead of explicitly estimating the values of $\{\beta_i\}_{i=1}^{\ell}$, we aim to ensure that the learned representation $\mathbf{F}$ respects the hierarchy of decoupled neighborhoods. Specifically, we require that the average intra-neighborhood similarity increases monotonically with the frequency level increasing. To this end, as shown in Figure 2, we derive a tractable surrogate objective $\mathcal{L}_{tra}$, which enforces the desired structural ordering through a ranking-based mechanism as

$$\mathcal{L}_{tra} = \mathcal{L}_{rank} + \mathcal{L}_{var} + \mathcal{L}_{neg}, \tag{4}$$

$$\text{with } \mathcal{L}_{rank} = \frac{16}{\ln(1+\mathrm{e}^2)} \sum_{u=1}^{n} \sum_{i=1}^{\ell} i|\mathcal{N}_u^{(i)}| \sum_{s=1}^{i} \ln\left(1+\mathrm{e}^{\mu_u^{(s-1)}-\mu_u^{(s)}}\right), \tag{5}$$

$$\mathcal{L}_{var} = \sum_{u=1}^{n} \sum_{i=1}^{\ell} \sum_{v\in\mathcal{N}_u^{(i)}} \left(\mu_u^{(i)} - (\mathbf{f}_u)^\top \mathbf{f}_v\right)^2, \tag{6}$$

$$\mathcal{L}_{neg} = \sum_{u=1}^{n} \left(1 + \frac{4\sum_{i=1}^{\ell}|\mathcal{N}_u^{(i)}|}{|\mathcal{N}_u^{(0)}|}\right) \sum_{v\in\mathcal{N}_u^{(0)}} \left((\mathbf{f}_u)^\top \mathbf{f}_v\right)^2, \tag{7}$$

where $\mu_u^{(i)} = 1/|\mathcal{N}_u^{(i)}| \sum_{v\in\mathcal{N}_u^{(i)}} \mathbf{f}_u^\top \mathbf{f}_v$ represents the average similarity at level i for $i \in [\ell]$.

By minimizing the surrogate $\mathcal{L}_{tra}$, we optimize the representation space $\mathbf{F}$ by the decoupled neighborhoods. Specifically, the term $\mathcal{L}_{rank}$ is a soft ranking loss that forces sample u to be closer to neighbors from high-frequency decoupled graphs than to those from lower-frequency ones; The term $\mathcal{L}_{var}$ is a variance penalty that enforces the consistency of pairwise similarities within the same decoupled neighborhood, ensuring structural consistency for edges with equal reliability; The term $\mathcal{L}_{neg}$ is a negative constraint that suppresses the similarities between samples and their disconnected ones. Thus, by constraining the pairwise similarities between edge frequency levels, $\mathcal{L}_{rank}$ effectively acts as a proxy for the implicit weights $\{\beta_i\}_{i=1}^{\ell}$ without requiring their exact values. In the following theorem, we indicate that the tractable surrogate objective is the upper bound of structural alignment objective.

Theorem 1. *It holds that* $\mathcal{L}_{alig} \leq \mathcal{L}_{tra} + 2\sum_{i=1}^{\ell} \beta_i^2 \mathrm{vol}_{\mathbf{A}^{(i)}}$.

Our method adopts a mini-batch training strategy to deal with large-scale datasets. In each iteration, graph construction and loss evaluation are performed solely on a sampled mini-batch rather than the entire dataset, enabling scalable training on large-scale datasets. Finally, the detailed procedure of our proposed DNi-MMC method is summarized in Algorithm 1 shown in the Appendix.

4.3 Algorithm Analysis

We now establish a theoretical guarantee for our framework. Given the cluster indicator vector $\mathbf{y}$, the cluster centroid of the j-th cluster for $j \in [k]$ in the joint representation space can be formulated as $\mathbf{u}_j = 1/|\mathcal{C}_j| \sum_{u\in\mathcal{C}_j} \mathbf{f}_u$. Then, we provide an upper bound on the k-means cost $\mathcal{J} = \sum_{j=1}^{k} \sum_{u\in\mathcal{C}_j} \|\mathbf{f}_u - \mathbf{u}_j\|_2^2$ to justify the minimization of $\mathcal{L}_{tra}$.

Theorem 2. *The k-means cost on the representation $\mathbf{F}$ is bounded by*

$$\mathcal{J} \leq n - k + \sqrt{k\left(\mathcal{L}_{tra} + 2\sum_{i=1}^{\ell} \beta_i^2 \mathrm{vol}_{\mathbf{A}^{(i)}}\right)} - \frac{\min_{i\in[\ell]}\{\mathrm{vol}_{\mathbf{A}^{(i)}}\}}{\max_{j\in[k]}\{|\mathcal{C}_j|\}} \sum_{i=1}^{\ell} \beta_i h_i. \tag{8}$$

Theorem 2 confirms that minimizing $\mathcal{L}_{tra}$ is able to compress the value of $\mathcal{J}$, leading to intra-cluster compactness of representation $\mathbf{F}$.

In this part, we further analyze our algorithm under a mini-batch training setting. Let $\mathcal{B}$ be sampled uniformly without replacement from $[n]$, with $|\mathcal{B}| = b$. Under the cluster indicator $\mathbf{y}$, there exist some cases $\mathcal{C}_j \cap \mathcal{B} = \emptyset$ for $k \in [k]$, making the number of clusters in the batch data less than k. We use $k_{\mathcal{B}}$ to denote the number of clusters in $\mathcal{B}$. Thus, we have

Corollary 1. *Suppose the size of cluster satisfies $|\mathcal{C}_j| \geq n/(ak)$ with a constant $a \geq 1$ for $j \in [k]$. Let $\mathcal{B}$ be sampled uniformly without replacement from $[n]$, with $|\mathcal{B}| = b > ak$. Then, the global k-means cost $\mathcal{J}$ satisfies*

$$\mathcal{J} \leq \frac{nb}{b-ak} + \frac{n}{b-ak} \cdot \mathbb{E}_{\mathcal{B}} \left[-k_{\mathcal{B}} - \frac{\min_{i\in[\ell]}\{\mathrm{vol}_{\tilde{\mathbf{A}}^{(i)}}\}}{\max_{j\in[k]}\{|\mathcal{C}_j \cap \mathcal{B}|\}} \sum_{i=1}^{\ell} \tilde{\beta}_i \tilde{h}_i + \sqrt{k_{\mathcal{B}} \left(\tilde{\mathcal{L}}_{tra} + 2 \sum_{i=1}^{\ell} \tilde{\beta}_i{}^2 \mathrm{vol}_{\tilde{\mathbf{A}}^{(i)}} \right)} \right],$$

where $\mathbf{A}^{(i)}$, $\mathcal{L}_{tra}$, h_i are defined on the data in $\mathcal{B}$, and $k_{\mathcal{B}}$ is the number of clusters on the data in $\mathcal{B}$.

It should be noticed that $\tilde{\mathbf{A}}^{(i)}$, $\tilde{\mathcal{L}}_{tra}$, $\tilde{h}_i$ are defined on the data in $\mathcal{B}$ by the same computational manners with $\mathbf{A}^{(i)}$, $\mathcal{L}_{tra}$, h_i.

Corollary 1 confirms that the mini-batch training with minimizing $\tilde{\mathcal{L}}_{tra}$ also ensure the intra-cluster compactness of representation $\mathbf{F}$.

5 Experiments

In this section, we first conduct a series of experiments to study the effectiveness and superiority of our method on multiple multi-modal datasets. Then, we perform experiments to analyze the effectiveness of diverse components and different configurations of the proposed method. Finally, we apply our method to the multi-modal attribute graph clustering task and evaluate its effectiveness.

5.1 Experimental Settings

Datasets. We employ five multi-modal datasets in the experiments. These include CUB [1], Youtube Wolf et al. (2011), nuswide Zhen et al. (2019), AWA [2], Reuters Amini et al. (2009). The Reuters dataset contains high-dimensional features, and in our experiments, we reduce its dimensionality by retaining the 300 columns with the largest variances, resulting in a 300-dimensional input for this dataset. For all datasets, feature values are normalized to the range $[0, 1]$ before training.

Evaluation Metrics. We adopt four metrics to evaluate clustering performance, including accuracy (ACC), normalized mutual information (NMI), adjusted Rand index (ARI), and F-score. For all measures, larger values indicate better clustering results. To measure inter-modal conflicts, we define modality conflict rate (MCR) as the proportion of mismatched cluster labels among ℓ modalities as

$$\mathrm{MCR} = \frac{1}{n\ell(\ell-1)} \sum_{i=1}^{\ell} \sum_{j\neq i}^{\ell} \sum_{u=1}^{n} \mathbb{1}(p_u^{(i)} \neq p_u^{(j)}), \tag{9}$$

where $p_u^{(i)}$ is the cluster indicator for sample u by performing k-means on modality i. Cluster indicators are aligned across modalities prior to calculation.

Implementation Details. In our method, the modality-specific extractor and feature fusion module are implemented several fully connected layers. The entire framework is implemented in PyTorch Paszke et al. (2019) and trained using the Adam optimizer Kingma (2014) with a learning rate of -01. We train the model for 2,500 epochs on each dataset. The number of nearest neighbors ζ is searched via grid search over $\{10, 15, 20, 25, 30, 35, 45, 50\}$. Each configuration of the proposed method is repeated multiple times to ensure accurate reporting of mean clustering performance. All experiments are conducted on an Ubuntu 20.04 system equipped with a single NVIDIA RTX 5090 GPU.

5.2 Clustering Performance Evaluation

In this part, we compare the proposed method with 12 different multi-modal clustering methods, including three shallow methods (AWP Nie et al. (2018), EEOMVC Wang et al. (2023a),

[1] https://www.vision.caltech.edu/visipedia/CUB-200.html

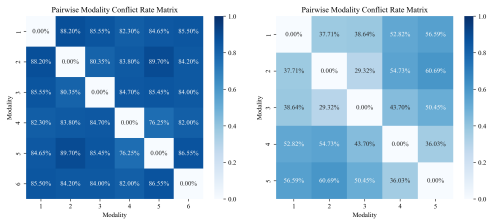
[2] https://cvml.ist.ac.at/AwA/

CAMVC Zhang et al. (2024)) and nine deep methods (DealMVC Yang et al. (2023), CVCL Chen et al. (2023), SCMVC Wu et al. (2024), DDMVC Chen et al. (2025b), MVCAN Xu et al. (2024), DRCFMVC Li et al. (2026c), DMVCS Li et al. (2026a), and BMvC Li et al. (2026b)) to verify its superiority.

Table 1: The multi-modal clustering performance measured by ACC, NMI, ARI, and Fscore of all compared methods on diverse datasets. The highest and the second-highest values under each metric are **red** and teal green, respectively. MCR denotes the modality conflict rate.

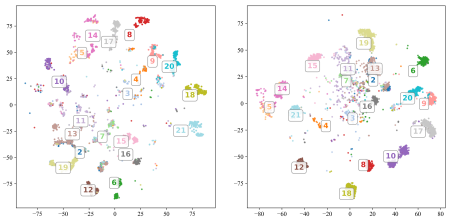
Datasets	Methods	AWP	EEOMVC	CAMVC	DealMVC	CVCL	SCMVC	DDMVC	MVCAN	DRCFMVC	DMVCS	GMAE	BMvC	Ours	Improv.
CUB	ACC	81.17	65.00	81.66	61.00	79.33	72.83	74.33	80.67	67.37	61.65	71.50	86.67	**89.97**	3.3
	NMI	75.20	66.49	79.67	66.10	71.03	68.71	68.25	78.42	68.54	68.65	75.55	82.21	**84.50**	2.29
	ARI	65.52	51.32	69.65	51.27	61.54	55.82	55.70	67.77	55.13	52.60	61.63	75.53	**79.67**	4.14
	Fscore	69.00	56.43	72.73	58.04	66.17	61.66	61.42	72.99	61.87	57.94	69.00	78.44	**82.41**	3.97
Youtube	ACC	31.40	26.25	25.70	36.60	46.15	41.60	54.40	29.75	43.17	27.80	44.35	47.67	**61.31**	6.91
	NMI	18.26	11.80	11.58	23.98	31.68	29.72	42.03	22.31	31.24	18.15	31.16	34.25	**45.57**	3.54
	ARI	12.13	7.08	6.60	15.88	23.32	20.27	33.45	13.64	22.42	10.63	22.80	26.23	**39.55**	6.1
	Fscore	22.02	17.27	16.13	26.24	31.50	29.08	41.56	25.35	30.51	20.47	31.31	34.70	**46.27**	4.71
nuswide	ACC	41.82	57.01	63.93	56.47	57.99	48.71	63.43	59.12	54.38	49.29	60.75	**68.68**	68.64	-0.04
	NMI	39.98	41.65	49.80	41.25	43.22	37.40	46.81	44.30	40.16	40.96	48.99	54.29	**55.55**	1.26
	ARI	22.13	34.54	44.19	35.69	38.22	26.62	42.48	38.12	33.05	31.79	41.39	50.18	**50.77**	0.59
	Fscore	33.44	41.28	49.83	42.47	44.13	36.99	48.86	44.56	40.57	38.95	48.48	55.08	**56.15**	1.07
AWA	ACC	63.61	61.99	64.43	38.78	35.76	38.15	36.21	56.43	36.77	54.75	-	67.16	**69.51**	2.35
	NMI	70.85	68.91	73.44	50.35	49.15	51.35	52.84	68.24	51.13	70.11	-	75.97	**75.99**	0.02
	ARI	52.91	52.08	56.33	28.23	23.46	28.46	27.94	47.68	28.36	43.99	-	59.17	**61.91**	2.74
	Fscore	54.03	53.17	57.31	28.67	26.13	28.65	32.47	48.79	28.99	45.34	-	60.89	**61.81**	0.92
Reuters	ACC	30.78	27.85	50.20	33.26	44.02	39.92	50.14	52.29	43.56	38.57	-	50.36	**55.86**	3.57
	NMI	12.57	10.92	28.52	16.07	25.61	20.98	**37.14**	30.35	27.45	27.71	-	27.65	36.97	-0.17
	ARI	0.39	-0.93	23.79	7.93	19.12	13.59	29.01	28.21	18.84	19.85	-	26.24	**31.05**	2.04
	Fscore	34.31	33.59	40.81	30.47	35.13	33.23	43.46	43.31	37.94	36.08	-	39.57	**46.39**	2.93

Clustering Performance Comparison. The clustering performance in terms of four different evaluation metrics of all compared methods on five datasets and MCR of each dataset are reported in Table 1. In addition, we provide the pairwise modality conflicts rate for Youtube and Reuters datasets in Figure 3. Our experimental results suggest that our proposed method delivers consistently better clustering performance than all compared approaches across all datasets with respect to most evaluation metrics. Taking the AWA dataset as an example, under 70.65% modality conflict rate, our proposed method surpasses the second-best one (BMvC) by approximately 2.35, and 2.74 percentage points in terms of ACC and ARI, respectively. These performance gains clearly demonstrate the advantages of our approach. By leveraging decoupled neighborhoods to encode relative consistent structures, the method is able to mitigate the cross modality conflicts and learn more discriminative representations, thereby yielding improved multi-modal clustering results. Compared with deep MMC approaches based on weighted feature fusion, contrastive learning objectives, and collaborative training strategies, the proposed method also achieves superior clustering performance. A common limitation of these existing methods is their inability to effectively address modality inconsistency. In contrast, our neighborhood decoupling mechanism separates the mixed structural information across different modalities into components with higher to lower consistency with respect to the underlying cluster structures, enabling the model to more effectively preserve neighborhood relationships.



(a) Youtube (b) Reuters

Figure 3: Visualization of the pairwise modality conflict rates.



(a) BMvC (b) Ours

Figure 4: Visualization of representations obtained from the BMvC method and our algorithm on AWA dataset.

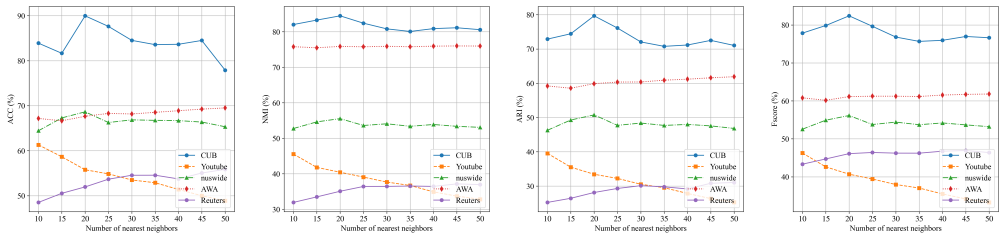
Visualization of Modality-joint Representation. To provide a more intuitive comparison with the learned modality-joint representation, we visualize the representation distributions of the first 20 clusters obtained by our method and BMvC approach on AWA dataset via the t-SNE approach

in Figure 4. The experimental results suggest that our method obtains clearer cluster structures compared with the BMvC approach, further demonstrating the effectiveness of our method.

5.3 Ablation Studies

In this part, we conduct a series of experiments to analyze the effectiveness of diverse components and different configurations in the proposed method, and our findings are reported below.

Hyper-parameter Sensitivity Analysis. The number of nearest neighbors is a critical hyper-parameter in the proposed method. To examine its sensitivity, we evaluate the clustering performance in terms of ACC, NMI, ARI, and Fscore of the proposed method across five datasets by varying the number of nearest neighbors within the range $\{10, 15, 20, 25, 30, 35, 40, 45\}$, as shown in Fig. 5. The results demonstrate that the performance of the proposed method remains relatively stable across different numbers of nearest neighbors. When the number of nearest neighbors is set to 20, the proposed method is able to achieve promising performance on most of the datasets. Thus, we suggest setting the number of nearest neighbors to 20 when the proposed method is applied to other unseen datasets from some real-world applications.



(a) ACC (b) NMI (c) ARI (d) Fscore

Figure 5: The clustering performance of the proposed method on different datasets under different number of nearest neighbors with respect to ACC, NMI, ARI, and Fscore.

Table 2: The performance of our proposed method with different variants of loss functions on three datasets with respect to ACC, NMI, ARI, and Fscore.

Losses	Youtube				AWA				Reuters			
	ACC	NMI	ARI	Fscore	ACC	NMI	ARI	Fscore	ACC	NMI	ARI	Fscore
$\mathcal{L}_{align}$	40.92	24.16	18.14	26.68	65.33	74.34	57.62	58.69	51.05	30.98	26.79	42.12
$\mathcal{L}_{rank}$	41.56	28.39	20.91	28.94	39.19	56.71	30.24	32.65	54.26	36.35	**35.15**	45.58
$\mathcal{L}_{rank} + \mathcal{L}_{neg}$	59.90	45.16	38.51	45.50	68.74	75.91	60.53	61.63	53.48	35.37	28.24	45.75
$\mathcal{L}_{var} + \mathcal{L}_{neg}$	43.86	30.35	22.54	31.52	67.71	74.66	59.71	60.37	46.28	28.11	20.41	41.39
$\mathcal{L}_{rank} + \mathcal{L}_{var} + \mathcal{L}_{neg}$	**61.31**	**45.57**	**39.55**	**46.27**	**69.51**	**75.99**	**61.91**	**61.81**	**55.86**	**36.97**	31.05	**46.39**

Effectiveness of Loss Function. Our optimization objective consists of three parts, i.e., the soft rank loss $\mathcal{L}_{rank}$, the variance penalty loss $\mathcal{L}_{var}$, and the negative constraint loss $\mathcal{L}_{neg}$. To verify the effectiveness of each part, we report the clustering performance in terms of ACC, NMI, ARI, and Fscore of our method under different combinations of three parts on three datasets in Table 2. We also provide the clustering performance of our method under the structure alignment loss $\mathcal{L}_{alig}$ in Table 2. From the results, we observe that our method reaches its best performance only when equipped with all three parts of the proposed optimization objective.

5.4 Application for Multi-modal Attributed Graph Clustering

The multi-modal attributed graph is an important data structure that captures complex interconnections among entities associated with attributes from multiple modalities, such as text and images. Multi-modal attributed graph clustering reveals latent cluster structures within these node relationships, benefiting a wide range of applications, including medical analysis Zheng et al. (2022) and recommendation systems Sun et al. (2020). However, due to the diverse sources of information, the

inherent inconsistency across modalities poses significant challenges for effectively leveraging both graph topology and node attributes for clustering.

Clustering Performance Comparison. To study the effectiveness and superiority of the proposed method on the multi-modal attributed graph clustering task, we conduct experiments on datasets from the multi-modal attributed graph benchmark Yan et al. (2025), including Grocery-S, Photo from Amazon Ni et al. (2019), and Reddit-S from RedCaps Desai et al. (2021). Each node in these graphs contains textual and visual attributes. Following the strong baseline settings reported in the benchmark, we adopt features extracted using Gemma Team et al. (2024) for the textual modality and CLIP ViT Radford et al. (2021) for the visual modality. A total of 11 different methods are employed in the experiments, including two attribute graphs clustering methods (MVGRL Hassani & Khasahmadi (2020), and Dink-Net Liu et al. (2023)), five multi-modal clustering methods (EMVGC-LG Wen et al. (2023), GLSEF Wang et al. (2023b), FastMICE Huang et al. (2023a), AEVC Liu et al. (2024), and BMvC Li et al. (2026b)), and four multi-modal attributed graph clustering methods (MCGC Pan & Kang (2021), MGCCN Liu et al. (2022), LMGEC Fettal et al. (2023), and VGMGC Chen et al. (2025a)).

Table 3: The multi-modal attribute graph clustering performance measured by ACC, NMI, ARI, and Fscore of all compared methods on diverse datasets. The highest and the second-highest values under each metric are **red** and teal green, respectively.

Dataset	Metric	MVGRL	Dink-Net	EMVGC-LG	GLSEF	FastMICE	AEVC	MCGC	MGCN	LMGE	VGMC	BMvC	Ours	Improv.
Grocery-S	ACC	39.48	21.64	43.31	20.22	29.05	38.69	17.80	32.56	33.87	25.15	45.22	**47.32**	2.1
	NMI	37.56	13.99	**51.57**	12.34	24.71	41.36	8.98	29.30	34.23	21.19	48.90	48.87	-2.7
	Fscore	28.90	17.32	33.06	14.33	19.21	30.08	8.43	26.66	30.36	23.56	32.03	**34.59**	1.53
	ARI	21.14	3.24	27.61	4.20	13.80	25.51	1.04	18.11	18.63	9.25	38.82	**40.12**	1.3
Reddit-S	ACC	71.02	21.98	75.34	20.55	75.58	61.19	9.59	71.44	64.89	60.72	77.52	**81.04**	3.52
	NMI	76.52	20.78	78.07	22.31	77.51	60.76	4.28	74.55	66.84	65.65	**84.12**	82.89	-1.23
	Fscore	63.01	19.56	71.44	14.90	71.73	50.54	3.04	63.40	54.27	53.34	**78.50**	77.98	-0.52
	ARI	65.96	9.09	69.63	5.32	69.95	47.50	-0.08	66.05	58.02	51.16	76.82	**78.31**	1.49
Photo	ACC	36.44	33.58	50.78	24.12	5-	37.79	43.45	OOM	49.44	OOM	55.56	**56.36**	0.8
	NMI	26.80	22.43	50.19	1.91	46.72	34.93	10.52	OOM	46.08	OOM	53.52	**54.26**	0.74
	Fscore	21.60	32.23	32.82	23.15	33.96	25.51	14.45	OOM	46.91	OOM	52.81	**53.17**	0.36
	ARI	7.08	6.94	23.32	-0.59	24.33	14.07	2.82	OOM	22.74	OOM	**28.64**	28.37	-0.27

The multi-modal attribute graph can be represented by a triplet $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \{\mathbf{X}^{(s)}\}_{i=1}^{\ell}\}$, where $\mathcal{V}$ denotes a set of nodes, $\mathcal{E}$ is a set of edges connecting nodes in $\mathcal{V}$, $\{\mathbf{X}^{(i)}\}_{i=1}^{\ell}$ is a set of multi-modal attributes. To make the proposed method tolerant to the multi-modal attribute graph clustering task, we capture the modality-joint feature representation from multi-modal attributes $\{\mathbf{X}^{(i)}\}_{i=1}^{\ell}$. In addition, we construct multi-modal graphs from corresponding attributes and obtain the decoupled neighborhoods from them and $\mathcal{E}$, ensuring our learning loss is leveraged to guide the learning process of modality-joint feature representation.

The clustering results of all methods on the three datasets in terms of ACC, NMI, ARI, and F-score are reported in Tab. 3. As shown, the proposed method consistently outperforms other approaches across all datasets and most of the evaluation metrics. For instance, on Grocery-S dataset, our method exceeds the second-best performance by approximately 2.1%, 1.53%, and 1.3% in terms of ACC, ARI, and F-score, respectively. These substantial performance gains demonstrate the effectiveness of the proposed method for multi-modal attributed graph clustering.

6 Conclusions

In this paper we propose a novel multi-modal clustering method to mitigate the effects of modality inconsistency on learning modality-joint representations. The method adopts a neighborhood decoupling strategy to decompose the multi-modal graphs into edge-disjoint ones based on edge frequencies, from which we introduce a structure alignment learning loss to guide the learning process of modality-joint representations. A theoretical analysis shows that our method equipped with this loss function is able to obtain compact cluster structures. Extensive experimental results demonstrate the effectiveness of the proposed method. Since the current neighborhood decoupling mechanism primarily relies on the completeness of multi-modal data, our future work will focus on extending the method to handle incomplete scenarios, such as those involving missing modalities.

REFERENCES

Massih R Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. *Advances in Neural Information Processing Systems*, 22, 2009.

Jianpeng Chen, Yawen Ling, Jie Xu, Yazhou Ren, Shudong Huang, Xiaorong Pu, Zhifeng Hao, Philip S Yu, and Lifang He. Variational graph generator for multiview graph clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2025a.

Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16752–16761, 2023.

Zhe Chen, Xiao-Jun Wu, Tianyang Xu, Hui Li, and Josef Kittler. Deep discriminative multi-view clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025b.

Chenhang Cui, Yazhou Ren, Jingyu Pu, Jiawei Li, Xiaorong Pu, Tianyi Wu, Yutao Shi, and Lifang He. A novel approach for effective multi-view clustering with information-theoretic perspective. *Advances in Neural Information Processing Systems*, 36:44847–44859, 2023.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

Chakib Fettal, Lazhar Labiod, and Mohamed Nadif. Simultaneous linear multi-view attributed graph representation learning and clustering. In *ACM International Conference on Web Search and Data Mining*, pp. 303–311, 2023.

Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024.

Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, pp. 4116–4126. PMLR, 2020.

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

Shudong Huang, Ivor W Tsang, Zenglin Xu, and Jiancheng Lv. Measuring diversity in graph learning: A unified framework for structured multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5869–5883, 2021.

Shudong Huang, Yixi Liu, Ivor W Tsang, Zenglin Xu, and Jiancheng Lv. Multi-view subspace clustering by joint measuring of consistency and diversity. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8270–8281, 2022.

Weitian Huang, Sirui Yang, and Hongmin Cai. Generalized information-theoretic multi-view clustering. *Advances in Neural Information Processing Systems*, 36:58752–58764, 2023b.

Zhao Kang, Wangtao Zhou, Zhitong Zhao, Junming Shao, Meng Han, and Zenglin Xu. Large-scale multi-view subspace clustering in linear time. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 4412–4419, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

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

Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.

Junfeng Li, Zhangyu Xu, Nanpei Li, Kaijun Zhang, Guangyong Xiong, Minjie Sun, Chao Hou, Jingjing Ji, Fan Zhang, Junwen Zhong, et al. Ai-embodied multi-modal flexible electronic robots with programmable sensing, actuating and self-learning. *Nature Communications*, 16(1):8818, 2025.

Pengyuan Li, Dongxia Chang, Yiming Wang, Zisen Kong, Linhua Kong, and Yao Zhao. Disentangled contrastive multi-view clustering via semantic relevance invariance. *IEEE Transactions on Knowledge and Data Engineering*, 2026a.

Zhenglai Li, Jun Wang, Chang Tang, Xinzhong Zhu, Wei Zhang, and Xinwang Liu. Balanced multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–16, 2026b. doi: 10.1109/TPAMI.2026.3688728.

Ziqi Li, Junjie Liao, Yonghong Zhang, and Tianyang Xu. Dual-reliable contrastive fusion for multi-view clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026c.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11174–11183, 2021.

Liang Liu, Zhao Kang, Jiajia Ruan, and Xixu He. Multilayer graph contrastive clustering network. *Information Sciences*, 613:256–267, 2022.

Suyuan Liu, Ke Liang, Zhibin Dong, Siwei Wang, Xihong Yang, Sihang Zhou, En Zhu, and Xinwang Liu. Learn from view correlation: An anchor enhancement strategy for multi-view clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 26151–26161, 2024.

Yue Liu, Ke Liang, Jun Xia, Sihang Zhou, Xihong Yang, Xinwang Liu, and Stan Z Li. Dink-net: Neural clustering on large graphs. In *International conference on machine learning*, pp. 21794–21812. PMLR, 2023.

Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. Simple contrastive multi-view clustering with data-level fusion. In *International Joint Conference on Artificial Intelligence*, pp. 4697–4705, 2024.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pp. 188–197, 2019.

Feiping Nie, Lai Tian, and Xuelong Li. Multiview clustering via adaptively weighted procrustes. In *International Conference on Knowledge Discovery and Data Mining*, pp. 2022–2030, 2018.

Feiping Nie, Wei Chang, Rong Wang, and Xuelong Li. Learning an optimal bipartite graph for subspace clustering via constrained laplacian rank. *IEEE Transactions on Cybernetics*, 53(2): 1235–1247, 2021.

Erlin Pan and Zhao Kang. Multi-view contrastive graph clustering. *Advances in neural information processing systems*, 34:2148–2159, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. Multi-modal knowledge graphs for recommender systems. In *ACM International Conference on Information and Knowledge Management*, pp. 1405–1414, 2020.

Chuan Tang, Miaomiao Li, Jun Wang, Chang Tang, Jiahe Jiang, Tianyi Wang, En Zhu, and Xinwang Liu. Multi-view clustering via high-order bipartite graph learning and tensor low-rank representation. *IEEE Transactions on Knowledge and Data Engineering*, 37(11):6521–6534, 2025. doi: 10.1109/TKDE.2025.3603594.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Daniel J Trosten, Sigurd Løkse, Robert Jenssen, and Michael C Kampffmeyer. On the effects of self-supervision and contrastive alignment in deep multi-view clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 23976–23985, 2023.

Fangdi Wang, Jiaqi Jin, Jingtao Hu, Suyuan Liu, Xihong Yang, Siwei Wang, Xinwang Liu, and En Zhu. Evaluate then cooperate: Shapley-based view cooperation enhancement for multi-view clustering. In *Advances in Neural Information Processing Systems*, 2024.

Hao Wang, Yan Yang, and Bing Liu. Gmc: Graph-based multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1116–1129, 2020.

Jun Wang, Chang Tang, Zhiguo Wan, Wei Zhang, Kun Sun, and Albert Y Zomaya. Efficient and effective one-step multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2023a.

Penglei Wang, Danyang Wu, Rong Wang, and Feiping Nie. Multi-view graph clustering via efficient global-local spectral embedding fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3268–3276, 2023b.

Yang Wang, Lin Wu, Xuemin Lin, and Junbin Gao. Multiview spectral clustering via structured low-rank matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems*, 29 (10):4833–4843, 2018.

Yi Wen, Suyuan Liu, Xinhang Wan, Siwei Wang, Ke Liang, Xinwang Liu, Xihong Yang, and Pei Zhang. Efficient multi-view graph clustering with local and global structure preservation. In *ACM International Conference on Multimedia*, pp. 3021–3030, 2023.

Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 529–534. IEEE, 2011.

Song Wu, Yan Zheng, Yazhou Ren, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, and Lifang He. Self-weighted contrastive fusion for deep multi-view clustering. *IEEE Transactions on Multimedia*, 2024.

Yanwanyu Xi, Chang Tang, Jun-Jie Huang, Xingchen Hu, Yuanyuan Liu, and Xinwang Liu. Contrastive and dual adversarial representation learning for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 38(1):170–182, 2026. doi: 10.1109/TKDE.2025.3611368.

Teng Xiao, Huaisheng Zhu, Zhengyu Chen, and Suhang Wang. Simple and asymmetric graph contrastive learning without augmentations. *Advances in neural information processing systems*, 36:16129–16152, 2023.

Teng Xiao, Huaisheng Zhu, Zhiwei Zhang, Zhimeng Guo, Charu C Aggarwal, Suhang Wang, and Vasant G Honavar. Efficient contrastive learning for fast and accurate inference on graphs. In *Forty-first International Conference on Machine Learning*, 2024.

Jie Xu, Yazhou Ren, Guofeng Li, Lili Pan, Ce Zhu, and Zenglin Xu. Deep embedded multi-view clustering with collaborative training. *Information Sciences*, 573:279–290, 2021.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16051–16060, 2022.

Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, and Xiaofeng Zhu. Investigating and mitigating the side effects of noisy views for self-supervised clustering algorithms in practical multi-view scenarios. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 22957–22966, 2024.

Hao Yan, Chaozhuo Li, Jun Yin, Zhigang Yu, Weihao Han, Mingzheng Li, Zhengxin Zeng, Hao Sun, and Senzhang Wang. When graph meets multimodal: benchmarking and meditating on multimodal attributed graph learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5842–5853, 2025.

Xihong Yang, Jin Jiaqi, Siwei Wang, Ke Liang, Yue Liu, Yi Wen, Suyuan Liu, Sihang Zhou, Xinwang Liu, and En Zhu. Dealmvc: Dual contrastive calibration for multi-view clustering. In *ACM International Conference on Multimedia*, pp. 337–346, 2023.

Chao Zhang, Xiuyi Jia, Zechao Li, Chunlin Chen, and Huaxiong Li. Learning cluster-wise anchors for multi-view clustering. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 16696–16704, 2024.

Hengrui Zhang, Qitian Wu, Yu Wang, Shaofeng Zhang, Junchi Yan, and Philip S Yu. Localized contrastive learning on graphs. *arXiv preprint arXiv:2212.04604*, 2022.

Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10394–10403, 2019.

Shuai Zheng, Zhenfeng Zhu, Zhizhe Liu, Zhenyu Guo, Yang Liu, Yuchen Yang, and Yao Zhao. Multi-modal graph learning for disease prediction. *IEEE Transactions on Medical Imaging*, 41(9): 2207–2216, 2022.

Yanran Zhu, Xiao He, Chang Tang, Xinwang Liu, Yuanyuan Liu, and Kunlun He. Multi-view adaptive fusion network for spatially resolved transcriptomics data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):8889–8900, 2024. doi: 10.1109/TKDE.2024.3450333.

Xin Zou, Ruimeng Liu, Chang Tang, Zhenglai Li, Xinwang Liu, Kunlun He, and Wanqing Li. Learning disentangled representations for generalized multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–16, 2026. doi: 10.1109/TPAMI.2026.3687339.

APPENDIX

Appendix Contents

A Necessary statement **15**

A.1 AI use statement 15

A.2 Ethics statement 15

A.3 Reproducibility statement 15

B Additional Details **15**

B.1 Algorithm Details 15

B.2 Detailed Information of Datasets 15

B.3 Scalability Analysis 15

B.4 Additional Ablation Studies 16

C Omitted Details of Proof **17**

C.1 Proof of Theorem 1 17

C.2 Proof of Theorem 2 19

C.3 Proof of Corollary 1 20

A NECESSARY STATEMENT

A.1 AI USE STATEMENT

In this work, we used generative AI tools to assist with presentation refinement, code debugging and refactoring. Generative AI tools were also used to improve the clarity and organization of some technical explanations. All AI-assisted content was carefully reviewed by the authors. In particular, AI-assisted code modifications were inspected and tested before being used in the experiments. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

A.2 ETHICS STATEMENT

This work focuses on the theoretical and algorithmic study of multi-modal clustering and does not involve human-subject experiments or the collection of personally identifiable information.

A.3 REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate the reproducibility of our results. The assumptions and formal definitions underlying the theoretical analysis are stated explicitly in the main text, while complete proofs of the theoretical results are provided in the appendix. Detailed descriptions of the proposed algorithm are also provided in the main paper and appendix.

B ADDITIONAL DETAILS

In this part, we present additional details about our algorithm, information of the datasets, and additional experimental results.

B.1 ALGORITHM DETAILS

In this paper, we give the detailed procedure of our proposed DNi-MMC method in Algorithm 1.

Algorithm 1: DNi-MMC.

Input: DNi-MMC Model ϕ, dataset $\{\mathbf{X}^{(i)}\}_{i=1}^{\ell}$, number of clusters k, number of nearest neighbors ζ, and maximum training epoch T;
Output: clustering results $\mathcal{A}_1, \ldots, \mathcal{A}_k$;
for $t = 1, ..., T$ **do**
  Sampling mini-batch of data, construct modality-specific graphs $\{\tilde{\mathbf{W}}^{(i)}\}_{i=1}^{\ell}$ and obtain decoupled graphs $\{\tilde{\mathbf{A}}^{(i)}\}_{i=1}^{\ell}$;
  Extract modality-joint representation from the mini-batch of data;
  Calculate loss $\tilde{\mathcal{L}}_{tra}$;
  Do backpropagation and update parameters of the DNi-MMC Model ϕ;
end
Obtain the complete modality-joint representation $\mathbf{F}$;
Perform k-means clustering algorithm on the learned modality-joint representation $\mathbf{F}$ to obtain clustering results $\mathcal{A}_1, \ldots, \mathcal{A}_k$.

B.2 DETAILED INFORMATION OF DATASETS

The detailed information about these datasets used in our experiments is listed in Tab. 4 and Tab. 5.

B.3 SCALABILITY ANALYSIS

Our method adopts a mini-batch training strategy to deal with large-scale datasets. In each iteration, graph construction and loss evaluation are performed solely on a sampled mini-batch rather than the

Table 4: Detailed information about five datasets.

Datasets	# samples	# clusters	# features
CUB	600	10	4,096; 300
Youtube	2,000	10	2,000; 1,024; 64; 512; 64; 647
nuswide	9,000	10	4,096; 300
AWA	10,158	50	4,096; 4,096
Reuters	18,758	6	300; 300; 300; 300

Table 5: Detailed information about four multi-modal attribute graph datasets.

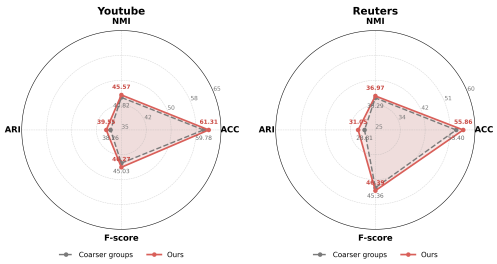
Datasets	# nodes	# edges	# clusters	# attributes
Movies	16,672	160,802	20	3,072; 768
Toys	20,695	113,402	18	3,072; 768
Reddit-S	15,894	283,080	20	3,072; 768
Photo	36,731	615,958	12	3,072; 768

entire dataset, enabling scalable training on large-scale datasets. We further report the total training time and GPU memory usage of deep clustering methods on the Reuters dataset in Tab. 6. Our method achieves the lowest training time among the compared deep clustering methods, although its peak GPU memory consumption is the second highest because multiple modality-specific neighborhood structures and frequency groups are maintained during training.

Table 6: The training time consumption and GPU Memory usage for different methods on the Reuters dataset.

Methods	DealMVC	CVCL	SCMVC	DDMVC	MVCAN	DRCFMVC	DMVCS	BMvC	Ours
Training time (s)	546.98	855.14	314.71	934.55	1269.87	421.38	665.48	907.53	**309.14**
GPU Memory cost (MB)	501.06	687.82	**323.83**	3368	1774	1116.84	675.94	11117.4	5627.03 (2552.47 for $\mathcal{N}_u^{(0)}$)

B.4 Additional Ablation Studies



(a) ACC (b) NMI

Figure 6: The clustering performance of the proposed method on two datasets under coarser group-wise neighborhood decoupling and ours.

In this part, we study the effectiveness of the fine-grained decoupling strategy. To this end, we divide the neighborhoods across different modalities into three frequency-based groups, named coarser groups as follows:

- High-frequency neighborhoods: Edges appearing in more than half of the modalities;
- Low-frequency neighborhoods: Edges appearing in at least one but no more than half of the modalities;
- Non-neighborhoods: Edges appearing in no modalities.

Note that replacing our fine-grained neighborhood decoupling strategy with this coarser stratification only affects datasets with more than two modalities. Therefore, we report the comparison results on the YouTube and Reuters datasets in Fig. 6. Our proposed method consistently outperforms the coarser

grouping baseline, confirming the necessity and effectiveness of our fine-grained neighborhood decoupling design.

C Omitted Details of Proof

C.1 Proof of Theorem 1

In this part, we present the details for proving Theorem 1.

Proof. For $i \in [\ell]$ and $u \in [n]$, we rewrite the term $\left(\beta_i - (\mathbf{f}_u)^\top \mathbf{f}_v\right)^2$ by adding an intermediate term $\mu_u^{(i)} = 1/|\mathcal{N}_u^{(i)}| \sum_{v \in \mathcal{N}_u^{(i)}} (\mathbf{f}_u)^\top \mathbf{f}_v$ denoting the local neighborhoods similarity as

$$\begin{aligned}\left(\beta_i - (\mathbf{f}_u)^\top \mathbf{f}_v\right)^2 &= \left[\left(\beta_i - \mu_u^{(i)}\right) + \left(\mu_u^{(i)} - (\mathbf{f}_u)^\top \mathbf{f}_v\right)\right]^2 \\ &= \left(\beta_i - \mu_u^{(i)}\right)^2 + \left(\mu_u^{(i)} - (\mathbf{f}_u)^\top \mathbf{f}_v\right)^2 + 2\left(\beta_i - \mu_u^{(i)}\right)\left(\mu_u^{(i)} - (\mathbf{f}_u)^\top \mathbf{f}_v\right).\end{aligned}$$

Summing over all $v \in \mathcal{N}_u^{(i)}$, we have

$$\begin{aligned}\sum_{v \in \mathcal{N}_u^{(i)}} \left(\beta_i - (\mathbf{f}_u)^\top \mathbf{f}_v\right)^2 = &\sum_{v \in \mathcal{N}_u^{(i)}} \left(\beta_i - \mu_u^{(i)}\right)^2 + \sum_{v \in \mathcal{N}_u^{(i)}} \left(\mu_u^{(i)} - (\mathbf{f}_u)^\top \mathbf{f}_v\right)^2 \\ &+ 2\left(\beta_i - \mu_u^{(i)}\right) \sum_{v \in \mathcal{N}_u^{(i)}} \left(\mu_u^{(i)} - (\mathbf{f}_u)^\top \mathbf{f}_v\right).\end{aligned}$$

According to the definition of $\mu_u^{(i)}$, we have

$$\sum_{v \in \mathcal{N}_u^{(i)}} \left(\mu_u^{(i)} - (\mathbf{f}_u)^\top \mathbf{f}_v\right) = 0.$$

Thus, the last term of the above equality vanishes, leaving

$$\sum_{v \in \mathcal{N}_u^{(i)}} \left(\beta_i - (\mathbf{f}_u)^\top \mathbf{f}_v\right)^2 = \sum_{v \in \mathcal{N}_u^{(i)}} \left(\beta_i - \mu_u^{(i)}\right)^2 + \sum_{v \in \mathcal{N}_u^{(i)}} \left(\mu_u^{(i)} - (\mathbf{f}_u)^\top \mathbf{f}_v\right)^2.$$

Afterthat, we define $\mu_u^{(0)} = 1/\left|\mathcal{N}_u^{(0)}\right| \sum_{v \in \mathcal{N}_u^{(0)}} (\mathbf{f}_u)^\top \mathbf{f}_v$ as the average similarity over the non-neighborhood of sample u. Thus, $\mu_u^{(0)}, \mu_u^{(i)} \in [-1, 1]$. Then, We bound $\left(\beta_i - \mu_u^{(i)}\right)^2$ by considering different cases.

Case 1: We first consider $\mu_u^{(0)} - \mu_u^{(i)} \geq 0$. By $(x + y + z)^2 \leq 2x^2 + 4y^2 + 4z^2$, we have

$$\left(\beta_i - \mu_u^{(i)}\right)^2 = \left(\beta_i - \mu_u^{(0)} + \left(\mu_u^{(0)} - \mu_u^{(i)}\right)\right)^2 \leq 2\beta_i^2 + 4\left(\mu_u^{(0)}\right)^2 + 4\left(\mu_u^{(0)} - \mu_u^{(i)}\right)^2.$$

Because $\mu_u^{(0)} - \mu_u^{(i)} \in [0, 2]$, it follows that

$$\left(\mu_u^{(0)} - \mu_u^{(i)}\right)^2 \leq 2\left(\mu_u^{(0)} - \mu_u^{(i)}\right) \leq \frac{4}{\ln(1 + \mathrm{e}^2)} \ln\left(1 + \mathrm{e}^{\mu_u^{(0)} - \mu_u^{(i)}}\right),$$

where the last inequality follows by that for any $x \in [0, 2]$, here exist a constant $C_1 = 2/\ln(1 + \mathrm{e}^2)$ such that $x \leq C_1 \ln(1 + \mathrm{e}^x)$ holds.

Thus, we have

$$\left(\beta_i - \mu_u^{(i)}\right)^2 \leq 2\beta_i^2 + 4\left(\mu_u^{(0)}\right)^2 + \frac{16}{\ln(1 + \mathrm{e}^2)} \ln\left(1 + \mathrm{e}^{\mu_u^{(0)} - \mu_u^{(i)}}\right).$$

Case 2: We then consider $\mu_u^{(0)} - \mu_u^{(i)} < 0$ and $\mu_u^{(i)} \leq \beta_i$. Then, we have $0 \leq \beta_i - \mu_u^{(i)} \leq \beta_i - \mu_u^{(0)}$, which gives that

$$\left(\beta_i - \mu_u^{(i)}\right)^2 \leq \left(\beta_i - \mu_u^{(0)}\right)^2 \leq 2\beta_i^2 + 2\left(\mu_u^{(0)}\right)^2 \leq 2\beta_i^2 + 4\left(\mu_u^{(0)}\right)^2 + \frac{16}{\ln(1+\mathrm{e}^2)} \ln\left(1 + \mathrm{e}^{\mu_u^{(0)} - \mu_u^{(i)}}\right).$$

Case 3: We then consider $\mu_u^{(0)} - \mu_u^{(i)} < 0$, $\mu_u^{(i)} > \beta_i$ and $\mu_u^{(i)} \leq 2|\mu_u^{(0)}|$. This case gives that

$$\left(\beta_i - \mu_u^{(i)}\right)^2 \leq 4\left(\mu_u^{(0)}\right)^2 \leq 2\beta_i^2 + 4\left(\mu_u^{(0)}\right)^2 + \frac{16}{\ln(1+\mathrm{e}^2)} \ln\left(1 + \mathrm{e}^{\mu_u^{(0)} - \mu_u^{(i)}}\right).$$

Case 4: We then consider $\mu_u^{(0)} - \mu_u^{(i)} < 0$, $\mu_u^{(i)} > \beta_i$ and $\mu_u^{(i)} > 2|\mu_u^{(0)}|$. It holds that

$$\mu_u^{(0)} - \mu_u^{(i)} \geq -|\mu_u^{(0)}| - \mu_u^{(i)} > -\frac{3}{2}\mu_u^{(i)} \geq -\frac{3}{2}.$$

Thus, we have

$$\left(\beta_i - \mu_u^{(i)}\right)^2 \leq \left(\mu_u^{(i)}\right)^2 \leq 1 \leq \frac{16}{\ln(1+\mathrm{e}^2)} \ln\left(1 + \mathrm{e}^{-3/2}\right) \leq \frac{16}{\ln(1+\mathrm{e}^2)} \ln\left(1 + \mathrm{e}^{\mu_u^{(0)} - \mu_u^{(i)}}\right),$$

Combining the above cases, we obtain

$$\left(\beta_i - \mu_u^{(i)}\right)^2 \leq 2\beta_i^2 + 4\left(\mu_u^{(0)}\right)^2 + \frac{16}{\ln(1+\mathrm{e}^2)} \ln\left(1 + \mathrm{e}^{\mu_u^{(0)} - \mu_u^{(i)}}\right).$$

Next, we rewrite $\mu_u^{(0)} - \mu_u^{(i)}$ as

$$\mu_u^{(0)} - \mu_u^{(i)} = \sum_{t=1}^{i} \left(\mu_u^{(t-1)} - \mu_u^{(t)}\right).$$

For any $x, y \in \mathbb{R}$, it holds that $1 + \mathrm{e}^{x+y} \leq (1+\mathrm{e}^x)(1+\mathrm{e}^y)$, which gives

$$\ln\left(1 + \mathrm{e}^{\mu_u^{(0)} - \mu_u^{(i)}}\right) \leq \sum_{t=1}^{i} \ln\left(1 + \mathrm{e}^{\mu_u^{(t-1)} - \mu_u^{(t)}}\right) \leq i \sum_{t=1}^{i} \ln\left(1 + \mathrm{e}^{\mu_u^{(t-1)} - \mu_u^{(t)}}\right).$$

Therefore, we have

$$\left(\beta_i - \mu_u^{(i)}\right)^2 \leq 2\beta_i^2 + 4\left(\mu_i^{(0)}\right)^2 + \frac{16i}{\ln(1+\mathrm{e}^2)} \sum_{t=1}^{i} \ln\left(1 + \mathrm{e}^{\mu_u^{(t-1)} - \mu_u^{(t)}}\right).$$

Taking all decoupled graphs into consideration, we have

$$\begin{aligned}
&\sum_{u=1}^{n}\sum_{i=1}^{\ell}\sum_{v\in\mathcal{N}_u^{(i)}} \left(\beta_i - \left(\mathbf{f}_u\right)^\top \mathbf{f}_v\right)^2 \\
&\leq 2\sum_{u=1}^{n}\sum_{i=1}^{\ell} |\mathcal{N}_u^{(i)}|\beta_i^2 + 4\sum_{i=1}^{\ell}\sum_{u=1}^{n} |\mathcal{N}_u^{(i)}| \left(\mu_i^{(0)}\right)^2 \\
&\quad + \frac{16}{\ln(1+\mathrm{e}^2)} \sum_{u=1}^{n}\sum_{i=1}^{\ell} i|\mathcal{N}_u^{(i)}| \sum_{t=1}^{i} \ln\left(1 + \mathrm{e}^{\mu_u^{(t-1)} - \mu_u^{(t)}}\right) + \sum_{u=1}^{n}\sum_{i=1}^{\ell}\sum_{v\in\mathcal{N}_u^{(i)}} \left(\mu_u^{(i)} - \left(\mathbf{f}_u\right)^\top \mathbf{f}_v\right)^2 \\
&= 2\sum_{i=1}^{\ell} \mathrm{vol}_{\mathbf{A}^{(i)}}\beta_i^2 + 4\sum_{i=1}^{\ell}\sum_{u=1}^{n} |\mathcal{N}_u^{(i)}| \left(\mu_i^{(0)}\right)^2 \\
&\quad + \frac{16}{\ln(1+\mathrm{e}^2)} \sum_{u=1}^{n}\sum_{i=1}^{\ell} i|\mathcal{N}_u^{(i)}| \sum_{t=1}^{i} \ln\left(1 + \mathrm{e}^{\mu_u^{(t-1)} - \mu_u^{(t)}}\right) + \sum_{u=1}^{n}\sum_{i=1}^{\ell}\sum_{v\in\mathcal{N}_u^{(i)}} \left(\mu_u^{(i)} - \left(\mathbf{f}_u\right)^\top \mathbf{f}_v\right)^2.
\end{aligned}$$

By Jensen inequality, it holds that $\left(\mu_i^{(0)}\right)^2 \leq 1/|\mathcal{N}_i^{(0)}| \sum_{v\in\mathcal{N}_i^{(0)}} \left((\mathbf{f}_u)^\top \mathbf{f}_v\right)^2$. This gives

$$\begin{aligned}
&\sum_{i=1}^{\ell}\sum_{u=1}^{n}\sum_{v\in\mathcal{N}_u^{(i)}}\left(\beta_i - (\mathbf{f}_u)^\top \mathbf{f}_v\right)^2 + \sum_{u=1}^{n}\sum_{v\in\mathcal{N}_i^{(0)}}\left(\mathbf{f}_i^\top \mathbf{f}_v\right)^2 \\
&\leq 2\sum_{i=1}^{\ell}\beta_i^2 \mathrm{vol}_{\mathbf{A}^{(i)}} + \frac{16}{\ln(1+\mathrm{e}^2)}\sum_{u=1}^{n}\sum_{i=1}^{\ell} i|\mathcal{N}_u^{(i)}| \sum_{t=1}^{i}\ln\left(1+\mathrm{e}^{\mu_u^{(t-1)}-\mu_u^{(t)}}\right) \\
&\quad + \sum_{u=1}^{n}\sum_{i=1}^{\ell}\sum_{v\in\mathcal{N}_u^{(i)}}\left(\mu_u^{(i)} - (\mathbf{f}_u)^\top \mathbf{f}_v\right)^2 + \sum_{u=1}^{n}\left(1+\frac{4\sum_{i=1}^{\ell}|\mathcal{N}_u^{(i)}|}{|\mathcal{N}_i^{(0)}|}\right)\sum_{v\in\mathcal{N}_i^{(0)}}\left((\mathbf{f}_u)^\top \mathbf{f}_v\right)^2.
\end{aligned}$$

Finally, we have

$$\mathcal{L}_{alig} \leq \mathcal{L}_{tra} + 2\sum_{i=1}^{\ell}\beta_i^2 \mathrm{vol}_{\mathbf{A}^{(i)}}.$$

This completes the proof. □

C.2 Proof of Theorem 2

In this part, we present the details of the proof of Theorem 2.

Proof. By the definition of k-means cost on $\mathbf{F}$, we have

$$\mathcal{J} = \sum_{j=1}^{k}\sum_{u\in\mathcal{C}_j}\|\mathbf{f}_u - \mathbf{u}_j\|_2^2 = \sum_{u=1}^{n}\|\mathbf{f}_u - \tilde{\mathbf{y}}_u\mathbf{U}\|_2^2 = \left\|\mathbf{F} - \mathbf{Y}\left(\mathbf{Y}^\top\mathbf{Y}\right)^{-1}\mathbf{Y}^\top\mathbf{F}\right\|_F^2,$$

where $\tilde{\mathbf{y}}_u$ the one-hot variant of y_u. The equality holds by the definition of center $\mathbf{u}_j = 1/|\mathcal{C}_j|\sum_{v\in\mathcal{C}_j}\mathbf{f}_v$ and center matrix $\mathbf{U} = \left(\mathbf{Y}^\top\mathbf{Y}\right)^{-1}\mathbf{Y}^\top\mathbf{F}$. Now, we rewrite the k-means cost as

$$\begin{aligned}
\sum_{j=1}^{k}\sum_{u\in\mathcal{C}_j}\|\mathbf{f}_u - \mathbf{u}_j\|_2^2 &= \mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\right) - 2\mathrm{Tr}\left(\mathbf{F}^\top\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{Y}\right)^{-1}\mathbf{Y}^\top\mathbf{F}\right) \\
&\quad + \mathrm{Tr}\left(\mathbf{F}^\top\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{Y}\right)^{-1}\mathbf{Y}^\top\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{Y}\right)^{-1}\mathbf{Y}^\top\mathbf{F}\right) \\
&= n - \mathrm{Tr}\left(\mathbf{Y}\left(\mathbf{Y}^\top\mathbf{Y}\right)^{-1}\mathbf{Y}^\top\mathbf{F}\mathbf{F}^\top\right).
\end{aligned}$$

Let $\mathbf{Q} = \mathbf{Y}\left(\mathbf{Y}^\top\mathbf{Y}\right)^{-1/2}$, we have

$$\begin{aligned}
&\sum_{j=1}^{k}\sum_{u\in\mathcal{C}_j}\|\mathbf{f}_u - \mathbf{u}_j\|_2^2 \\
&= n - \mathrm{Tr}\left(\mathbf{Q}\mathbf{Q}^\top\mathbf{F}\mathbf{F}^\top\right) \\
&= n - \mathrm{Tr}\left(\mathbf{Q}\mathbf{Q}^\top\left(\sum_{i=1}^{\ell}\beta_i\mathbf{A}^{(i)} + \mathbf{I}\right)\right) + \mathrm{Tr}\left(\mathbf{Q}\mathbf{Q}^\top\left(\sum_{i=1}^{\ell}\beta_i\mathbf{A}^{(i)} + \mathbf{I} - \mathbf{F}\mathbf{F}^\top\right)\right). \qquad (10)
\end{aligned}$$

We fist build the lower bound for the term $\mathrm{Tr}\left(\mathbf{Q}\mathbf{Q}^\top\left(\sum_{i=1}^{\ell}\beta_i\mathbf{A}^{(i)} + \mathbf{I}\right)\right)$ in Eq. equation 10. Recall the definition of h_s, we have

$$h_s = \frac{\sum_{u=1}^{n}\sum_{v\in\mathcal{N}_u^{(i)}}\mathbb{I}[y_u = y_v]}{\sum_{u=1}^{n}\left|\mathcal{N}_u^{(i)}\right|} = \frac{\sum_{j=1}^{k}\sum_{u\in\mathcal{C}_j}\sum_{v\in\mathcal{C}_j}\mathbf{A}_{i,v}^{(i)}}{\mathrm{vol}_{\mathbf{A}^{(i)}}}.$$

This gives

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{Q}\mathbf{Q}^\top\left(\sum_{i=1}^{\ell}\beta_i\mathbf{A}^{(i)}+\mathbf{I}\right)\right) &= \sum_{u=1}^{n}\frac{1}{|\mathcal{C}_{y_u}|}+\sum_{i=1}^{\ell}\beta_i\sum_{j=1}^{k}\frac{1}{|\mathcal{C}_j|}\sum_{u\in\mathcal{C}_j}\sum_{v\in\mathcal{C}_j}\mathbf{A}^{(i)}_{u,v}\\ &= k+\sum_{i=1}^{\ell}\beta_i\sum_{j=1}^{k}\frac{1}{|\mathcal{C}_j|}\sum_{u\in\mathcal{C}_j}\sum_{v\in\mathcal{C}_j}\mathbf{A}^{(i)}_{u,v}\\ &\geq k+\frac{\min_{i\in[m]}\{\mathrm{vol}_{\mathbf{A}^{(i)}}\}}{\max_{j\in[k]}\{|\mathcal{C}_j|\}}\sum_{i=1}^{\ell}\beta_i h_i,\end{aligned}$$

where the second equality holds by $\sum_{u=1}^{n}1/|\mathcal{C}_{y_u}|=\sum_{j=1}^{k}|\mathcal{C}_j|1/|\mathcal{C}_j|=k$.

Next we analyze the upper bound for the term $\mathrm{Tr}\left(\mathbf{Q}\mathbf{Q}^\top\left(\sum_{i=1}^{\ell}\beta_i\mathbf{A}^{(i)}+\mathbf{I}-\mathbf{F}\mathbf{F}^\top\right)\right)$ in Eq. equation 10. We have

$$\begin{aligned}&\mathrm{Tr}\left(\mathbf{Q}\mathbf{Q}^\top\left(\sum_{i=1}^{\ell}\beta_i\mathbf{A}^{(i)}+\mathbf{I}-\mathbf{F}\mathbf{F}^\top\right)\right)\\ &\leq \left\|\mathbf{Q}\mathbf{Q}^\top\right\|_F\left\|\left(\sum_{i=1}^{\ell}\beta_i\mathbf{A}^{(i)}+\mathbf{I}-\mathbf{F}\mathbf{F}^\top\right)\right\|_F\\ &= \sqrt{k}\left\|\left(\sum_{i=1}^{\ell}\beta_i\mathbf{A}^{(i)}+\mathbf{I}-\mathbf{F}\mathbf{F}^\top\right)\right\|_F\\ &= \sqrt{k}\sqrt{\sum_{u=1}^{n}\sum_{i=1}^{\ell}\sum_{v\in\mathcal{N}_u^{(i)}}(\beta_i-\mathbf{f}_u^\top\mathbf{f}_v)^2+\sum_{u=1}^{n}\sum_{v\in\mathcal{N}_u^{(0)}}(\mathbf{f}_u^\top\mathbf{f}_v)^2},\end{aligned}$$

where the first equation holds by the fact that $\mathbf{Q}\mathbf{Q}^\top$ is an orthogonal projection matrix with its rank equal to k. Combining the above inequality with Theorem 1, we have

$$\mathcal{J}\leq n-k-\frac{\min_{i\in[\ell]}\{\mathrm{vol}_{\mathbf{A}^{(i)}}\}}{\max_{j\in[k]}\{|\mathcal{C}_j|\}}\sum_{i=1}^{\ell}\beta_i h_i+\sqrt{k\left(\mathcal{L}_{tra}+2\sum_{i=1}^{\ell}\beta_i^2\mathrm{vol}_{\mathbf{A}^{(i)}}\right)}.$$

This completes the proof. □

C.3 Proof of Corollary 1

In this part, we present the details for proving Corollary 1.

Proof. For each cluster $j\in\mathcal{S}_{\mathcal{B}}$, we define the local batch center as $\mathbf{u}_{j,\mathcal{B}}=1/|\mathcal{C}_j\cap\mathcal{B}|\sum_{u\in\mathcal{C}_j\cap\mathcal{B}}\mathbf{f}_u$. For any sample $u\in\mathcal{C}_j\cap\mathcal{B}$ and $\mathcal{C}_j\cap\mathcal{B}\neq\emptyset$, we formulate the batch-level k-means cost on clusters in $\mathcal{B}$ is $\mathcal{J}_{\mathcal{B}}$ as

$$\mathcal{J}_{\mathcal{B}}=\sum_{j\in\mathcal{S}_{\mathcal{B}}}\sum_{u\in\mathcal{C}_j\cap\mathcal{B}}\|\mathbf{f}_u-\mathbf{u}_{j,\mathcal{B}}\|_2^2.$$

By Theorem 2, we have

$$\mathcal{J}_{\mathcal{B}}\leq b-k_{\mathcal{B}}-\frac{\min_{i\in[\ell]}\{\mathrm{vol}_{\tilde{\mathbf{A}}^{(i)}}\}}{\max_{j\in[k]}\{|\mathcal{C}_j\cap\mathcal{B}|\}}\sum_{i=1}^{\ell}\tilde{\beta}_i\tilde{h}_i+\sqrt{k_{\mathcal{B}}\left(\tilde{\mathcal{L}}_{tra}+2\sum_{i=1}^{\ell}\tilde{\beta}_i{}^2\mathrm{vol}_{\tilde{\mathbf{A}}^{(i)}}\right)}.$$

During the sampling procedure, every sample $u\in[n]$ is selected into $\mathcal{B}$ with b/n probability. There exist some cases $\mathcal{C}_j\cap\mathcal{B}=\emptyset$ for $k\in[k]$, making the number of clusters in the batch data is less than

k. We use $\mathcal{S}_\mathcal{B} = \{j \in [k] : \mathcal{C}_j \cap \mathcal{B} \neq \emptyset\}$ to denote the set of clusters in $\mathcal{B}$. Let $\mathbf{u}_j = 1/|\mathcal{C}_j| \sum_{u \in \mathcal{C}_j} \mathbf{f}_u$ be the center of cluster $\mathcal{C}_j$ constructed from the total dataset. Then, we have

$$\mathbb{E}_\mathcal{B}\left[\sum_{j=1}^{k} \sum_{u \in \mathcal{C}_j \cap \mathcal{B}} \|\mathbf{f}_u - \mathbf{u}_j\|_2^2\right] = \frac{b}{n} \sum_{j=1}^{k} \sum_{u \in \mathcal{C}_j} \|\mathbf{f}_u - \mathbf{u}_j\|_2^2 = \frac{b}{n}\mathcal{J}.$$

For any sample $u \in \mathcal{C}_j \cap \mathcal{B}$ and $\mathcal{C}_j \cap \mathcal{B} \neq \emptyset$, we have

$$\|\mathbf{f}_u - \mathbf{u}_j\|_2^2 = \|\mathbf{f}_u - \mathbf{u}_{j,\mathcal{B}}\|_2^2 + \|\mathbf{u}_{j,\mathcal{B}} - \mathbf{u}_j\|_2^2 + 2\langle \mathbf{f}_u - \mathbf{u}_{j,\mathcal{B}},\, \mathbf{u}_{j,\mathcal{B}} - \mathbf{u}_j \rangle.$$

Summing over $u \in \mathcal{C}_j \cap \mathcal{B}$ for $j \in \mathcal{S}_\mathcal{B}$, the cross-term vanishes, because $\sum_{u \in \mathcal{C}_j \cap \mathcal{B}} (\mathbf{f}_u - \mathbf{u}_{j,\mathcal{B}}) = \mathbf{0}$. Thus, we have

$$\sum_{j=1}^{k} \sum_{u \in \mathcal{C}_j \cap \mathcal{B}} \|\mathbf{f}_u - \mathbf{u}_j\|_2^2 = \sum_{j \in \mathcal{S}_\mathcal{B}} \sum_{u \in \mathcal{C}_j \cap \mathcal{B}} \|\mathbf{f}_u - \mathbf{u}_j\|_2^2 = \mathcal{J}_\mathcal{B} + \mathcal{D}_\mathcal{B},$$

where $\mathcal{J}_\mathcal{B} = \sum_{j \in \mathcal{S}_\mathcal{B}} \sum_{u \in \mathcal{C}_j \cap \mathcal{B}} \|\mathbf{f}_u - \mathbf{u}_{j,\mathcal{B}}\|_2^2$ is the local batch cost, and $\mathcal{D}_\mathcal{B} = \sum_{j \in \mathcal{S}_\mathcal{B}} |\mathcal{C}_j \cap \mathcal{B}| \cdot \|\mathbf{u}_{j,\mathcal{B}} - \mathbf{u}_j\|_2^2$ measures the center drift. We can further obtain

$$\frac{b}{n}\mathcal{J} = \mathbb{E}_\mathcal{B}\left[\mathcal{J}_\mathcal{B}\right] + \mathbb{E}_\mathcal{B}\left[\mathcal{D}_\mathcal{B}\right].$$

Next, we analyze $\mathbb{E}_\mathcal{B}[\mathcal{D}_\mathcal{B}]$. Let $\mathcal{J}_j = \sum_{u \in \mathcal{C}_j} \|\mathbf{f}_u - \mathbf{u}_j\|_2^2$ be the cost on j-the cluster for $j \in [k]$. Conditional on $|\mathcal{C}_j \cap \mathcal{B}| = s \geq 1$, $\mathbf{u}_{j,\mathcal{B}}$ is the sample mean from a finite population $\mathcal{C}_j$. By the property of finite-population variance, we have

$$\mathbb{E}\left[\|\mathbf{u}_{j,\mathcal{B}} - \mathbf{u}_j\|_2^2 \;\middle|\; |\mathcal{C}_j \cap \mathcal{B}| = s\right] = \left(\frac{|\mathcal{C}_j| - s}{|\mathcal{C}_j| - 1}\right) \frac{\mathcal{J}_j}{s\,|\mathcal{C}_j|}.$$

Multiplying both sides by s, we obtain

$$\mathbb{E}\left[s \cdot \|\mathbf{u}_{j,\mathcal{B}} - \mathbf{u}_j\|_2^2 \;\middle|\; |\mathcal{C}_j \cap \mathcal{B}| = s\right] = \left(\frac{|\mathcal{C}_j| - s}{|\mathcal{C}_j| - 1}\right) \frac{\mathcal{J}_j}{|\mathcal{C}_j|} \leq \frac{\mathcal{J}_j}{|\mathcal{C}_j|}.$$

If $s = 0$, the cluster is unrepresented and its contribution to $\mathcal{D}_\mathcal{B}$ is $0 \leq \mathcal{J}_j/|\mathcal{C}_j|$. Thus, the upper bound holds for all $s \geq 0$. By the Law of Total Expectation, summing over all k clusters yields

$$\mathbb{E}_\mathcal{B}[\mathcal{D}_\mathcal{B}] = \sum_{j=1}^{k} \mathbb{E}_\mathcal{B}\left[|\mathcal{C}_j \cap \mathcal{B}| \cdot \|\mathbf{u}_{j,\mathcal{B}} - \mathbf{u}_j\|_2^2\right] \leq \sum_{j=1}^{k} \frac{\mathcal{J}_j}{|\mathcal{C}_j|}.$$

Applying the cluster condition, e.g., $|\mathcal{C}_j| \geq n/(ak)$, we obtain

$$\mathbb{E}_\mathcal{B}[\mathcal{D}_\mathcal{B}] \leq \frac{ak}{n} \sum_{j=1}^{k} \mathcal{J}_j = \frac{ak}{n}\mathcal{J}.$$

Thus, we have

$$\frac{b}{n}\mathcal{J} = \mathbb{E}_\mathcal{B}[\mathcal{J}_\mathcal{B}] + \mathbb{E}_\mathcal{B}[\mathcal{D}_\mathcal{B}] \leq \mathbb{E}_\mathcal{B}[\mathcal{J}_\mathcal{B}] + \frac{ak}{n}\mathcal{J}.$$

Rearranging terms establishes

$$\mathcal{J} \leq \frac{n}{b - ak} \cdot \mathbb{E}_\mathcal{B}[\mathcal{J}_\mathcal{B}],$$

Thus,

$$\mathcal{J} \leq \frac{n}{b - ak} \cdot \mathbb{E}_\mathcal{B}\left[b - k_\mathcal{B} - \frac{\min_{i \in [\ell]}\{\mathrm{vol}_{\tilde{\mathbf{A}}^{(i)}}\}}{\max_{j \in [k]}\{|\mathcal{C}_j \cap \mathcal{B}|\}} \sum_{i=1}^{\ell} \tilde{\beta}_i \tilde{h}_i + \sqrt{k_\mathcal{B}\left(\tilde{\mathcal{L}}_{tra} + 2\sum_{i=1}^{\ell} \tilde{\beta}_i{}^2 \mathrm{vol}_{\tilde{\mathbf{A}}^{(i)}}\right)}\right]$$

which completes the proof. □