MULTI-SCALE DIFFUSION-GUIDED GRAPH LEARNING WITH POWER-SMOOTHING RANDOM WALK CONTRAST FOR MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the notable advances in graph-based deep multi-view clustering, existing approaches still suffer from three critical limitations: (1) relying on static graph structures and being unable to model the global semantic relationships across views; (2) contamination from false negative samples in contrastive learning frameworks; and (3) a fundamental trade-off between cross-view consistency and view-specific discrimination. To address these issues, we introduce **M**ulti-sc**A**le diffusio**N**-guided **G**raph learning with p**O**wer-smoothing random walk contrast (**MANGO**) for multi-view clustering, a unified framework that combines adaptive multi-scale diffusion, random walk-driven contrastive learning, and structure-aware view consistency modeling. Specifically, the multi-scale diffusion mechanism leverages local entropy guidance to dynamically fuse similarity matrices across different diffusion steps, thereby achieving joint modeling of fine-grained local structures and overall global semantics. Additionally, we introduce a random walk-based correction strategy that explores high-probability semantic paths to filter out false negative samples, and applies a β-power transformation to adaptively reweight contrastive targets, thereby reducing noise propagation. To further reconcile the consistency-specificity dilemma, the view consistency module enforces semantic alignment across views by sharing structural embeddings, ensuring consistent local structures while preserving heterogeneous features. Extensive experiments on 12 datasets demonstrate the superior performance of MANGO.

1 INTRODUCTION

Multi-view Clustering (MVC) aims to partition data samples into meaningful clusters by leveraging the consensus and complementary information across multiple views. By exploiting the synergistic relationships between diverse data representations, MVC facilitates the discovery of underlying structures in complex datasets, positioning it as a key research area for integrating heterogeneous information sources and revealing intrinsic patterns (Gao et al., 2015; Liu et al., 2018; 2020; Xu et al., 2022a; Yan et al., 2024). From the perspective of learning paradigms, existing MVC methods can be broadly categorized into traditional techniques and deep learning-based models (Fang et al., 2023). Among them, deep learning-based approaches have attracted increasing attention due to their strong capacity for modeling intricate data distributions and extracting highly expressive feature representations (Huang et al., 2023; Liu et al., 2024; Tang & Liu, 2022).

Deep multi-view clustering leverages the nonlinear mapping capabilities of deep neural networks to capture the distinctive semantics of each view while effectively integrating complementary information, enabling strong performance in complex data scenarios (Lin et al., 2023; Xu et al., 2023; Yang et al., 2023). Given the ability to explicitly model the topological relationships within the data, graph-based deep multi-view clustering (GDMVC) methods have garnered considerable attention. For example, Wen et al. (2024) proposed an adaptive hybrid graph filter that combines high- and low-frequency signals with fused multi-view embeddings to improve clustering performance on graphs. Ren et al. (2024) dynamically fuse weighted graphs using deep autoencoders and graph convolution, enabling efficient self-supervised deep multi-view clustering. Additionally, to accurately capture the affinity relationships between sample pairs, numerous contrastive learning-driven methods for graph structure refinement have been proposed Gao et al. (2024); Liu et al. (2023); Smith et al.

(2025). For example, Yu et al. (2025) proposed a multi-view deep subspace clustering method leveraging contrastive learning and Cauchy-Schwarz divergence for interactive representation and clustering optimization. Chen et al. (2023) introduced a cross-view contrastive learning model that learns view-invariant and robust representations by contrasting cluster assignments across views. Additionally, Wang et al. (2023) integrated triple contrastive learning at both the feature representation and graph structure layers to generate a consensus similarity graph with a clear clustering structure.

Despite the notable progress achieved by recent graph-based deep multi-view clustering methods, three fundamental technical challenges remain unresolved. First, the reliance on static graph structures imposes inherent limitations in capturing complex semantics across multiple views. Specifically, such methods only rely on the local neighborhood relationship between samples to calculate the similarity, ignoring the global semantic connection between views. This limitation leads to inevitable information loss and distortion, which makes it difficult for the model to accurately capture complex cross-view semantic associations. Second, the issue of negative sample contamination remains prominent in graph-based contrastive learning frameworks. When constructing negative pairs, the model may mistakenly treat semantically similar samples as negatives, introducing false contrastive signals. These errors can accumulate through gradient backpropagation, forming a positive feedback loop of "noisy optimization" that progressively degrades the quality of similarity measures and weakens the effectiveness of contrastive learning. Finally, the dilemma of balancing semantic consistency and modality specificity plagues multimodal alignment strategies. This trade-off can undermine the separability of clusters, as over-alignment harms the uniqueness of the modality, while under-alignment destroys the cross-view semantic correspondence.

To address the aforementioned limitations, this study proposes Multi-scAle diffusion-guided Graph learning with pOwer-smoothing random walk contrast (MANGO) model, which contains three technical innovations. First, we introduce an adaptive multi-scale diffusion mechanism. This module dynamically fuses similarity matrices from multiple diffusion steps based on local entropy information to build a more resilient and semantically expressive graph structure. By modeling on multi-scale topology, MANGO can capture local details between directly connected samples and global semantic connections between distant samples. Second, to address the challenge of negative sample contamination, random walk path sampling is introduced to dynamically correct the sample distribution of contrastive learning. This technique explores high-probability semantic paths to filter out false negative sample pairs, and is supplemented by β-power transformation to adaptively weight negative samples. The combined method reduces noise propagation and enhances the accuracy of graph similarity estimation through iterative refinement. Third, regarding the trade-off between consistency and specificity, our structure-aware cross-view contrastive learning mechanism achieves a dual goal: to enforce semantic consistency through shared structural embeddings, while retaining modality-specific discriminative features through a view-aware attention mechanism. This balance solves the problem of cluster boundary ambiguity by coordinating global semantic alignment and local modality uniqueness. The core contributions of this paper include the following three aspects:

- A multi-scale diffusion mechanism is proposed to break the performance bottleneck brought by the fixed diffusion step size, dynamically fuse the similarity information under different step sizes, take into account both local structure exploration and global semantic modeling, and realize the effective capture of multi-granularity structural information.

- A random walk correction method is designed to optimize the distribution of contrastive learning targets. The hybrid transfer matrix is constructed by combining the t-step transfer matrix and the unit matrix, and the weight of negative samples is adjusted through β-power transformation to form a more discriminative contrast target, which reduces the impact of erroneous negative samples.

- We design a structure-aware view consistency module that simultaneously promotes semantic alignment across views and preserves modality-specific discriminative features, thereby improving clustering quality in heterogeneous multi-view scenarios.

- Extensive experiments are conducted on 12 benchmark multi-view datasets of varying types and scales. The results verify the effectiveness of MANGO compared with several state-of-the-art multi-view clustering methods.

2 Related Work

Deep multi-view clustering (DMVC) methods can be broadly categorized into three paradigms based on how they handle inter-view relationships: joint methods, alignment-based methods, and other methods.

Joint methods integrate feature learning and clustering into unified objectives, leveraging cross-view collaboration to enhance representation quality. For instance, Li et al. (2021) jointly learned both view-specific and consensus graphs while adaptively assigning weights to obtain high-confidence clustering results. Xia et al. (2022) built a self-supervised framework based on Euler transformation and $\ell_{1,2}$-norm, integrating representation learning and clustering. Hu et al. (2023) enhanced feature-level alignment by incorporating cluster-level contrastive learning and dynamic weight learning to promote more consistent deep representations.

Alignment-based methods, in contrast, focus on mapping view-specific representations into a shared subspace to promote consistency. Early work by Hassani & Khasahmadi (2020) introduced a node-graph dual-granularity alignment framework to address cross-hierarchical redundancy. Building on this, Liu et al. (2022) mitigated representation collapse by reducing inter-view redundancy from both sample and feature perspectives. Chen et al. (2023) proposed a clustering-aware contrastive learning mechanism, which directly enforced semantic consistency across views. Trosten et al. (2023) identified negative sample bias in traditional contrastive alignment and developed a variational alignment model that maximized mutual information.

In addition, some approaches combine these strategies or address specific issues such as noise and incomplete views. For example, Luo et al. (2018) pioneered the combination of consistency constraints and view-specific modeling to establish a unified theoretical framework for multi-view subspace representation. Ke et al. (2021) built a full-process integrated framework of feature extraction-fusion-comparison-clustering, verifying the feasibility of multi-task joint optimization. Xu et al. (2021) introduced a common-specific variable dual-channel mechanism to separate multi-view shared clustering features from view-unique information.

3 Method

This section provides a detailed introduction to the proposed MANGO model, which primarily consists of four components: the self-expressive module, contrastive learning module, view consistency module, and adaptive diffusion module, as illustrated in Figure 1.

3.1 Self-expressive module

Given a multi-view dataset $\{\mathbf{X}^v \in \mathbb{R}^{n\times d_v}\}_{v=1}^m$, where m is the number of views, n denotes the number of samples, and $d_1, d_2, ..., d_m$ are the dimensionality of each view, we propose a self-expressive module to effectively integrate heterogeneous views. Specifically, for the v-th view, we first obtain its embedded representation through the encoder: $\mathbf{Z}^v = f^v(\mathbf{X}^v)$, where f^v represents the encoder of v-th view, and $\mathbf{Z}^v$ is the learned embedded representation. The core of the module is the encoder that achieves self-reconstruction through sparse combination of latent features. The reconstruction process is defined as $\hat{\mathbf{X}}^v = \mathbf{C}^{\mathbf{v}}\mathbf{Z}^v$, where $\mathbf{C}^{\mathbf{v}}$ is the sparse coefficient matrix obtained by weighted fusion of sparse matrices under each view. For the $v-$th view, the sparse adjacency matrix within the view is generated by filtering the cosine similarity of the sample pairs with an adaptive threshold b, ensuring that the reconstruction process captures local structural dependencies. Finally, the reconstruction loss is obtained:

$$\mathcal{L}_{rec} = \frac{1}{2}\sum_{v=1}^{m}\left\|\mathbf{X}^v - \hat{\mathbf{X}}^v\right\|_F^2 \tag{1}$$

In addition, to prevent overfitting, we use the hybrid regularization term following (You et al., 2016):

$$\mathcal{L}_{reg} = \sum_{v=1}^{m}\lambda\|\mathbf{C}^v\|_1 + \frac{1-\lambda}{2}\|\mathbf{C}^v\|_F^2 \tag{2}$$

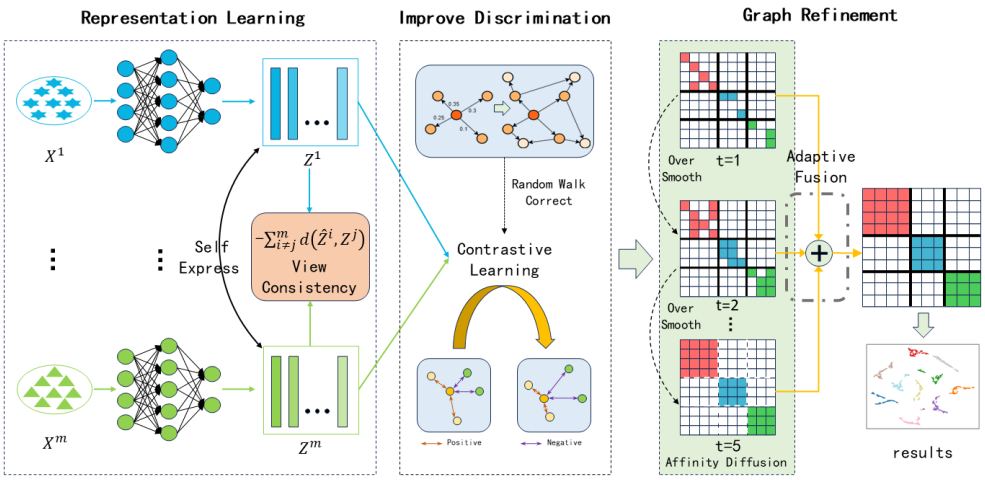


Figure 1: The Framework of MANGO. First, MANGO obtains embedding representations via view-specific MLP modules and captures data structure through reconstruction loss. Then, the view consistency module achieves semantic alignment by sharing structural embeddings, while contrastive learning and random walks filter high-probability semantic paths and eliminate pseudo-negative samples. Finally, the affinity matrix undergoes T steps of diffusion, and local structures and global semantics are fused through information entropy-based weighting.

where λ is used to balance the two regularization terms. This reconstruction loss, combined with hybrid regularization, yields embeddings that preserve both global semantics and local geometry, providing robust inputs for downstream modules.

3.2 Power-Smoothing Random Walk Enhanced Contrastive Learning

Contrastive learning is widely adopted for representation learning in unlabeled multi-view scenarios, where features are aligned by contrasting positive (similar) and negative (dissimilar) pairs. However, it relies on two strong assumptions: (1) positive pairs from different views of the same sample are semantically aligned, and (2) all negative pairs are unrelated. These assumptions often fail in practice—cross-view heterogeneity can make same-class instances appear dissimilar, producing false negatives (FNs), while the absence of labels makes it difficult to ensure true negatives. Such false negatives distort training signals, disrupt the manifold structure, and weaken the discriminative power of the learned representations.

To tackle this issue, we propose a power-smoothing random walk enhanced contrastive learning strategy, which integrates two key components: a random walk correction to capture high-order semantic relations, and a power-smoothing operation to reduce the impact of false negatives by refining similarity distributions.

Random walk-based correction mechanism: Traditional contrastive learning assumes equal importance for all non-anchor negative samples, overlooking the intrinsic structure of the data. The proposed random walk correction mechanism simulates random walks on the sample manifold to uncover and exploit this structural information, enabling more principled weighting of negative samples. Specifically, we first construct the affinity matrix $\mathbf{A}_{ij} = \exp(-\sigma \left\|\mathbf{z}_i - \mathbf{z}_j\right\|^2)$ through the Euclidean distance of sample embedding, where σ is the bandwidth parameter of the Gaussian kernel. After the affinity matrix $\mathbf{A}$ is constructed, it needs to be normalized and converted into a transfer matrix $\mathbf{M}$, where $\mathbf{M}_{ij}$ represents the one-step transfer probability from sample i to sample j.To better capture high-order manifold structures, we compute the t-step transition matrix $\mathbf{M}^t = \mathbf{M} \times \cdots \times \mathbf{M}$ (t times), where $\mathbf{M}$ is the one-step transition matrix.

$$\mathbf{M}_{ij} = \frac{\mathbf{A}_{ij}}{\sum_{k=1}^{n} \mathbf{A}_{ik}} \tag{3}$$

Finally, the interpolation parameter η is used to balance the self-connection strength and the manifold structure, formulated as $\mathbf{T} = \eta\mathbf{I} + (1-\eta)\mathbf{M}^t$, where $\mathbf{T}$ is the target distribution matrix, and $\mathbf{T}_{ij}$ denotes the degree to which sample j is a semantic neighbor of sample i. This value can be directly used as the negative sample weight in the intra-view contrastive loss.

Power-smoothing-induced contrastive learning : In addition, in order to enhance the robustness, we introduce a smoothing power operation on the basis of InfoNCE loss to control the overall strength of negative samples, which directly acts on the negative sample term in the contrast loss, thereby obtaining the expression of the intra-view contrast loss:

$$\mathcal{L}_{intra} = \frac{1}{m}\sum_{p=1}^{m}\left[-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp\left(\frac{s(\mathbf{z}_i^p,\mathbf{z}_i^p)}{\tau}\right)}{\exp\left(\frac{s(\mathbf{z}_i^p,\mathbf{z}_i^p)}{\tau}\right)+\left(\sum_{j\neq i}\mathbf{T}_{ij}\exp\left(\frac{s(\mathbf{z}_i^p,\mathbf{z}_j^p)}{\tau}\right)\right)^{\beta}}\right] \tag{4}$$

where n is the number of samples, $s(\mathbf{z_i}, \mathbf{z_j})$ represents the cosine similarity, and β is the power operation parameter. Compared with the standard InfoNCE loss, the negative sample item is subjected to the β power operation. This operation has a nonlinear smoothing effect on the negative sample item parameter, which is used to reduce the overall impact of negative samples, especially the impact of extreme value samples.

Similarly, the expression of the contrast loss between m views is as follows:

$$\mathcal{L}_{inter} = \frac{2}{m(m-1)}\sum_{p\neq q}\left[-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp\left(\frac{s^{(\mathbf{z}_i^p,\mathbf{z}_i^q)}}{\tau}\right)}{\exp\left(\frac{s^{(\mathbf{z}_i^p,\mathbf{z}_i^q)}}{\tau}\right)+\left(\sum_{j\neq i}\mathbf{W}_{ij}\exp\left(\frac{s^{(\mathbf{z}_i^p,\mathbf{z}_j^q)}}{\tau}\right)\right)^{\beta}}\right] \tag{5}$$

The difference is that $s(\mathbf{z}_i^p, \mathbf{z}_i^q)$ represents the cosine similarity between sample i in view p and sample j in view q, $\mathbf{z}_i^p$ is the embedding representation of sample i in view p, $\mathbf{z}_j^q$ is the embedding representation of sample j in view q, and $\mathbf{W}_{ij}$ is the uniform weight.

Finally, by integrating the aforementioned intra-view and inter-view contrastive losses, the Power-Smoothing Random Walk Enhanced Contrastive Learning framework can be formulated as follows, where μ is the balance parameter of the two contrast losses.

$$\mathcal{L}_{contra} = \mathcal{L}_{intra} + \mu\mathcal{L}_{inter} \tag{6}$$

3.3 View consistency module

Due to potential discrepancies in noise distributions and semantic emphasis across different views, embeddings of the same class can vary significantly among views. Directly inputting such misaligned embeddings into the subsequent fusion module will destroy the intrinsic consistency of the data and fuse the conflicting noise. Therefore, in the last step of representation learning, we introduced a view consistency module, which builds a mapping bridge between views to ensure that the representations from different views can be aligned and complement each other. Specifically, view consistency is to maximize the mutual information between representations of different views. Given two view embeddings $\mathbf{Z}^p$ and $\mathbf{Z}^q$, the mutual information is defined as:

$$I\left(\mathbf{Z}^p;\mathbf{Z}^q\right) = \iint p\left(\mathbf{Z}^p,\mathbf{Z}^q\right)\log\frac{p\left(\mathbf{Z}^p,\mathbf{Z}^q\right)}{p\left(\mathbf{Z}^p\right)p\left(\mathbf{Z}^q\right)}d\mathbf{Z}^p d\mathbf{Z}^q \tag{7}$$

$$I\left(\mathbf{Z}^p;\mathbf{Z}^q\right) \geq H\left(\mathbf{h}_i\right) - \mathbb{E}_{p(\mathbf{Z}^p,\mathbf{Z}^q)}\left[d\left(f_{p\to q}\left(\mathbf{Z}^p\right),\mathbf{Z}^q\right)\right] \tag{8}$$

where $d(\cdot,\cdot)$ represents cosine distance. The core of the view consistency module is to learn a mapping function f such that: $\hat{\mathbf{Z}}^p = f_{p\to q}(\mathbf{Z}^p) \approx \mathbf{Z}^q$. Therefore, the consistency loss function can be defined as $\mathcal{L}_{p\to q} = 1 - \frac{d(\hat{\mathbf{Z}}^p,\mathbf{Z}^q)}{\tau}$, where τ is used to control the sensitivity of the loss.

For the case of m views, we need to calculate the consistency loss between all view pairs:

$$\mathcal{L}_{consist} = \frac{1}{m(m-1)}\sum_{p\neq q}^{\mathbf{m}}\mathcal{L}_{p\to q} \tag{9}$$

3.4 ENTROPY-GUIDED MULTI-SCALE DIFFUSION FOR GRAPH REFINEMENT

After obtaining discriminative representations, a graph $\mathbf{A}$ is typically constructed to encode semantic similarities among samples, serving as a foundation for downstream clustering tasks. However, most existing graph-based approaches employ static graph structures, which are highly sensitive to the quality of the learned features. To mitigate this issue, we propose an entropy-guided multi-scale diffusion strategy for graph refinement, consisting of a multi-scale diffusion module and an entropy-guided learning module.

Multi-scale graph diffusion: Diffusion-based methods propagate information by repeatedly multiplying the transition probability matrix, enabling global semantic aggregation. Formally, the t-step diffusion is $\mathbf{A}^t = \mathbf{A} \times \cdots \times \mathbf{A}$, where $\mathbf{A}$ is the sparse similarity matrix, $\mathbf{A}_{ij}$ denotes the one-step transition probability from node i to j, and $\mathbf{A}^t$ represents the transition probability matrix after t steps. Unlike traditional diffusion, we compute matrices at multiple diffusion steps to capture structural information across scales, and dynamically fuse them to balance local structures and global semantics. Specifically, given the normalized affinity matrix $\mathbf{A}_{\text{norm}}$, we construct $\{\tilde{\mathbf{A}}^0, \tilde{\mathbf{A}}^1, \ldots, \tilde{\mathbf{A}}^t\}$, where $\tilde{\mathbf{A}}^0 = \mathbf{A}_{\text{norm}}$. After each step, we retain only the top-K elements per row and re-normalize to obtain $\tilde{\mathbf{A}}^t$.

Entropy-guided multi-scale graph learning: In this approach, entropy is used to evaluate the quality of the diffusion matrix by quantifying the uncertainty or uniformity in the distribution of connection weights. A lower entropy indicates a more concentrated distribution, which corresponds to more distinct and clearer semantic structures. Specifically, for each row $\tilde{\mathbf{A}}_i^t$ in the diffusion matrix $\tilde{\mathbf{A}}^t$, the entropy is computed over its non-zero elements, where $\tilde{\mathbf{A}}_{ij}^t$ denotes the weight of the connection between nodes i and j following the t-th diffusion step.

$$H(\tilde{\mathbf{A}}_i^t) = -\sum_{j:\tilde{\mathbf{A}}_{ij}^t>0} \tilde{\mathbf{A}}_{ij}^t \cdot \log \tilde{\mathbf{A}}_{ij}^t \tag{10}$$

Next, the average entropy of the matrix is computed as $\bar{H}(\tilde{\mathbf{A}}^t) = \frac{1}{n}\sum_{i=1}^{n} H(\tilde{\mathbf{A}}_i^t)$. The inverse of entropy is used as the weight of the scale because lower entropy indicates a more concentrated distribution, corresponding to a clearer category structure. This design enables our diffusion model to automatically adjust the weights and retain complementary information at multiple scales.

$$\mathbf{A}_{fusion} = \sum_{t=0}^{T} \frac{1}{\bar{H}(\tilde{\mathbf{A}}^t)} \tilde{\mathbf{A}}^t \tag{11}$$

To facilitate subsequent spectral clustering, we further apply symmetric normalization and diagonal enhancement to the final diffusion matrix. Specifically, the symmetric normalization is achieved by averaging the matrix with its transpose, while diagonal enhancement is performed by scaling the diagonal elements using an enhancement coefficient k.

$$\mathbf{A}_{final}[i,j] = \frac{1}{2}(\mathbf{A}_{fusion}[i,j] + \mathbf{A}_{fusion}[j,i]) \cdot k \tag{12}$$

3.5 THE OVERALL LOSS FUNCTION

Combining self-representation and regularization losses, by jointly random Walk modified power smoothing contrastive learning and view consistency modules, the overall loss function of our proposed MANGO is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{reg} + \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_{contra} + \gamma\mathcal{L}_{consist} \tag{13}$$

where hyper-parameters α, β, and γ balance the importance of the three terms.

4 EXPERIMENT

4.1 EXPERIMENTAL SETTINGS

Datasets: Twelve datasets with varying types and scales are used: Yale, ORL, BBC-Sport, Reuters, Scene-15, MSRC-v1, LandUse-21, Caltech101-20, ALOI-100, STL10, HandWritten, and MNIST-3V. More detailed descriptions of these datasets can be found in Table 1.

Baselines: We compare MANGO with eight SOTA multi-view clustering methods, including **MFLVC**(2022) (Xu et al., 2022b), **MSESC**(2023) (Cui et al., 2023), **CVCL**(2023) (Chen et al., 2023), **LSGMC**(2023) (Lan et al., 2023), **MVD**(2023) (Li et al., 2023), **DIVIDE**(2024) (Lu et al., 2024), **SCM**(2024) (Luo et al., 2024), **CANDY**(2024) (Guo et al., 2024).

Evaluation metrics: To comprehensively evaluate clustering performance, we adopt three widely used metrics: clustering accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (ARI), where higher values indicate better performance.

Table 1: The detail for experimental datasets

Dataset	Type	# Instances	# Classes	# Views
Yale	Face	165	15	3
ORL	Face	400	40	3
BBC-Sport	Text	544	5	2
Reuters	Text	1200	6	5
Scene-15	Scene	4485	15	3
MSRC-v1	Object	210	7	5
LandUse-21	Object	2100	21	3
Caltech101-20	Object	2386	20	6
ALOI-100	Object	10800	100	4
STL10	Object	13000	10	3
HandWritten	Digit	2000	10	6
MNIST-3V	Digit	60000	10	3

For fair comparison, the hyperparameters of all baseline methods are carefully tuned based on their publicly available code, and the best-performing settings are adopted. For the MANGO model, a three-layer MLP is employed to extract features for each view, with hidden layer sizes set to 1024, 512, 256. The input dimension corresponds to the original feature size, and the output dimension is fixed at 256. In all experiments, the bandwidth parameter of the Gaussian kernel σ is set to 0.3, the regularization parameter λ to 0.3, the interpolation parameter η to 1.2, the number of random walk steps t to 3, the temperature parameter τ to 0.6, and the contrastive loss weight μ to 0.1. Hyperparameters α, β, and γ are selected via grid search over the set 1e3,1e4, 1e4,1e5 and 1e5,1e6. Shallow learning experiments are implemented in MATLAB 2023b on a workstation with a 2.50GHz 7285H 32-core CPU and 128 GB RAM, while deep learning experiments are conducted using PyTorch 2.5.1 on an H20-NVLink GPU.

4.2 Comparison results

Table 2 records the experimental comparison of our proposed MANGO with other 8 comparison methods across twelve datasets, where the best and suboptimal performance are highlighted in bold and underline respectively, and the abbreviation "OM" indicates the occurrence of out of memory error. Through comprehensive experiments, we have the following observations:

1) Our MANGO model shows excellent clustering performance in all datasets, significantly outperforming its competitors in some scenarios. In particular, on the ALOI-100 dataset, our MANGO achieves 89.09% ACC, which is about 14.1% higher than the second-best algorithm CDMGC. These results suggest that MANGO effectively exploits the rich information among multi-view data by jointly leveraging the random walk-enhanced contrastive learning module and the view consistency mechanism.

2) Compared with existing contrastive learning-based algorithms (such as CANDY, DIVIDE, and SCM), the MANGO model achieves the best clustering performance in most cases. This demonstrates that our power-smoothing and random walk-enhanced contrastive learning mechanism effectively improves representation quality, which in turn leads to enhanced clustering results.

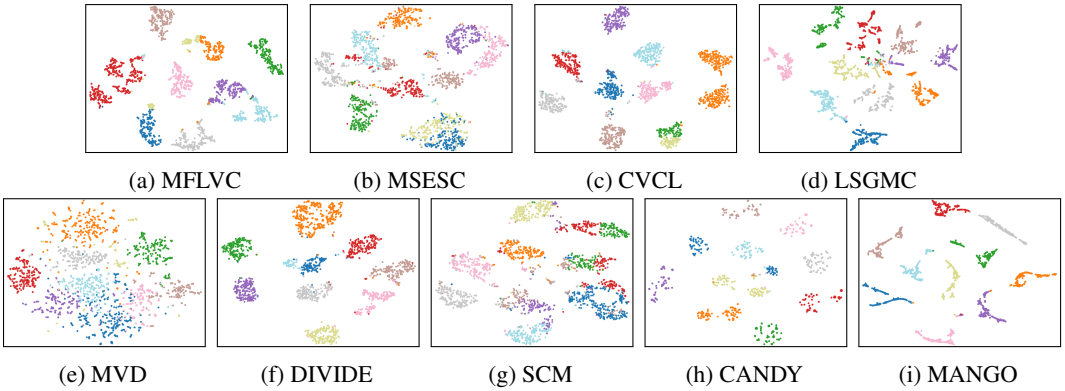
3) Shallow methods like LSGMC and MVD perform well on small datasets but struggle with scalability and often face out-of-memory errors on larger ones due to limited representational capacity. Deep methods such as SCM and CANDY improve on this by learning more expressive features, achieving better results on complex datasets like STL10 and MNIST-3V. Nonetheless, the proposed MANGO model consistently outperforms existing shallow and deep multi-view clustering methods in most cases, demonstrating its superior performance and robustness.

4) To further verify the effectiveness of our method, we take the HandWritten dataset as an example to visualize the clustering results of other MVC methods and our method. The t-SNE results are

Table 2: Clustering performance of all methods on twelve datasets

Metric	Dataset	MFLVC	MSESC	CVCL	LSGMC	MVD	DIVIDE	SCM	CANDY	MANGO
	ACC	0.5993	0.5455	0.6937	0.7152	0.6715	0.6182	0.5455	0.6333	**0.7163**
Yale	NMI	0.5772	0.5811	0.6783	0.7252	0.6999	0.6509	0.5708	0.6416	**0.7508**
	ARI	0.3375	0.3315	0.4939	0.5782	0.4366	0.4114	0.3305	0.4287	**0.5843**
	ACC	0.4325	0.7125	0.8173	0.8575	0.8897	0.7550	0.6575	0.6075	**0.9425**
ORL	NMI	0.5856	0.8116	0.8155	0.9434	0.9447	0.8783	0.8077	0.7542	**0.9663**
	ARI	0.2572	0.5356	0.5562	0.8202	0.8535	0.6844	0.6900	0.4578	**0.9108**
	ACC	0.7224	0.7757	0.6211	0.9412	0.7849	0.4467	0.7298	0.6728	**0.9650**
BBC-Sport	NMI	0.5344	0.6183	0.3645	0.8459	0.6828	0.1507	0.5570	0.3812	**0.8850**
	ARI	0.4937	0.5874	0.3137	0.8414	0.6216	0.1091	0.5052	0.3657	**0.9118**
	ACC	0.4216	0.4150	0.4696	0.3925	0.4641	0.5632	0.4883	0.5558	**0.5865**
Reuters	NMI	0.1895	0.2110	0.2605	0.2655	0.3512	0.3630	0.2642	0.2525	**0.3760**
	ARI	0.3210	0.1418	**0.4852**	0.1798	0.2712	0.2930	0.2258	0.2970	0.2872
	ACC	0.3138	0.4283	0.3719	0.4634	0.4130	0.4744	0.3485	0.3911	**0.4980**
Scene-15	NMI	0.3513	0.3989	0.3912	0.4647	0.3807	0.4845	0.3175	0.3597	**0.5042**
	ARI	0.1646	0.2438	0.3297	0.3240	0.2379	0.3071	0.1675	0.2149	**0.3388**
	ACC	0.8914	0.8114	0.9286	0.9181	0.8714	0.7381	0.6070	0.4738	**0.9495**
MSRC-v1	NMI	0.7804	0.6629	0.8721	0.8719	0.7506	0.6531	0.5613	0.3685	**0.9084**
	ARI	0.7389	0.5974	0.8626	0.8586	0.7116	0.5835	0.4742	0.2352	**0.8870**
	ACC	0.2495	0.2443	0.2922	0.3048	0.2575	0.3129	0.2440	0.2286	**0.3270**
LandUse-21	NMI	0.2663	0.3003	0.3339	0.3282	0.3308	0.2659	0.2942	0.2584	**0.3524**
	ARI	0.0966	0.0906	0.1355	0.1465	0.1144	0.1623	0.0984	0.0918	**0.1628**
	ACC	0.6266	0.4678	0.5107	0.6077	0.5731	0.6159	0.5771	0.4662	**0.6408**
Caltech101-20	NMI	0.7242	0.6710	0.4055	0.7222	0.6886	0.6266	0.4849	0.3461	**0.7329**
	ARI	0.5162	0.4580	0.5158	0.4983	0.4905	0.5044	**0.6271**	0.3168	0.5542
	ACC	0.6709	0.5579	0.7205	0.5323	0.7030	0.7499	0.6463	0.6963	**0.8909**
ALOI-100	NMI	0.7866	0.7506	0.6951	0.7177	0.8205	0.8288	0.7825	0.7060	**0.9114**
	ARI	0.5302	0.4562	0.5284	0.2307	0.5841	0.5870	0.5160	0.5821	**0.8047**
	ACC	0.1246	0.9331	0.7316	0.8309	0.1110	0.9174	0.9394	0.2802	**0.9677**
STL10	NMI	0.0360	0.8589	0.4966	0.8383	0.0013	0.8280	0.8598	0.0867	**0.9196**
	ARI	0.0080	0.8631	0.4930	0.7702	0.0000	0.8384	0.8707	0.0617	**0.9300**
	ACC	0.8990	0.9250	0.9104	0.9720	0.8708	0.8501	0.7940	0.9510	**0.9775**
HandWritten	NMI	0.8259	0.8540	0.8878	0.9381	0.9127	0.8277	0.7039	0.8796	**0.9480**
	ARI	0.7939	0.8400	0.8473	0.9497	0.8533	0.8086	0.6173	0.8691	**0.9505**
	ACC	0.9747	0.9407	0.8186	OM	OM	0.9840	0.9206	**0.9940**	0.9887
MNIST-3V	NMI	0.9405	0.9032	0.7803	OM	OM	0.9538	0.8545	**0.9667**	0.9663
	ARI	0.9443	0.8660	0.7104	OM	OM	0.9645	0.8388	0.9735	**0.9750**

shown in Figure 2. It can be seen that our method obtains more clear and compact clusters, which further confirms the superiority of our method.



(a) MFLVC (b) MSESC (c) CVCL (d) LSGMC

(e) MVD (f) DIVIDE (g) SCM (h) CANDY (i) MANGO

Figure 2: t-SNE visualization of the consensus affinity matrix on the HandWritten dataset

4.3 Parameter sensitivity analysis

This section studies the impact of three hyper-parameters α, β, and γ on the MANGO model. Specifically, we perform grid search by adjusting α, β, and γ in the set {1e3, 3e3, 5e3, 7e3, 9e3}, {1e4, 3e4, 5e4, 7e4, 9e4}, and {1e5, 3e5, 5e5, 7e5, 9e5} respectively. Figure 3 shows how the performance

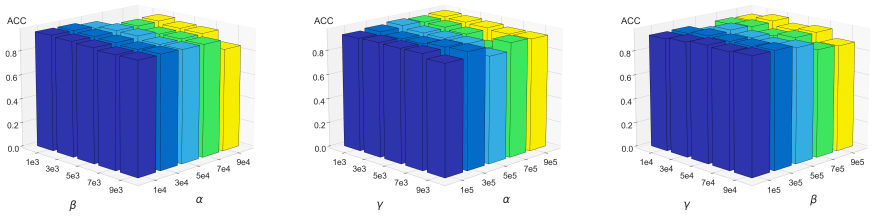


Figure 3: Parameters sensitivity analysis with parameters α, β, and γ on MSRC-v1.

of the model changes with various combinations of these parameters. The results indicate that the MANGO model performs well across the specified ranges of α, β, and γ, demonstrating robustness to variations in these hyperparameters.

4.4 ABLATION STUDY

Finally, we conduct comprehensive ablation experiments to evaluate the contribution of each module within the MANGO model. Specifically, we remove the contrastive loss ($\mathcal{L}_{contrat}$), view consistency loss ($\mathcal{L}_{consist}$), false negative (FN) adjustment strategy, and adaptive diffusion module from the complete MANGO model in various combinations and record the corresponding performance. The results on the MSRC-v1 and Reuters datasets are summarized in Table 3. It is evident that the full MANGO model achieves the best performance, demonstrating that the modules work synergistically to deliver superior clustering results. Moreover, the performance of configuration (c) surpasses that of (a), and (h) outperforms (g), indicating that the random walk-enhanced contrastive learning and view consistency modules effectively exploit view consistency, while the adaptive diffusion module efficiently captures the underlying graph structure.

Table 3: Ablation study on MSRC-v1 and Reuters dataset

	$\mathcal{L}_{rec}$	$\mathcal{L}_{contra}$	$\mathcal{L}_{consist}$	random	diffusion	MSRC-v1			Reuters		
						ACC	NMI	ARI	ACC	NMI	ARI
(a)	✓					0.770	0.758	0.662	0.502	0.347	0.267
(b)	✓	✓				0.795	0.780	0.699	0.510	0.347	0.269
(c)	✓	✓		✓		0.800	0.781	0.700	0.535	0.364	0.286
(d)	✓	✓		✓	✓	0.893	0.845	0.902	0.507	0.344	0.258
(e)	✓		✓			0.781	0.749	0.660	0.542	0.360	0.279
(f)	✓		✓		✓	0.790	0.770	0.681	0.546	0.360	0.270
(g)	✓	✓	✓	✓		0.863	0.818	0.748	0.551	0.340	0.261
(h)	✓	✓	✓	✓	✓	**0.950**	**0.908**	**0.887**	**0.587**	**0.376**	**0.287**

5 CONCLUSION

This paper proposes a novel deep multi-view clustering framework, which effectively learns a multi-view embedding representation with strong discriminative power by integrating the random walk modified contrastive learning module and the view consistency module. Among them, the random walk modified contrastive learning module enhances the adaptability of the model to complex data distribution by dynamically adjusting the weights of negative samples; the view consistency module realizes deep alignment across view feature spaces through a bidirectional projection mechanism. In addition, the introduction of the adaptive diffusion module can dynamically capture the multi-scale structural information of the data, effectively avoiding the over-smoothing and information loss problems commonly seen in traditional methods. Extensive experiments fully verify the superiority and effectiveness of MANGO.

ETHICS STATEMENT

In this study, we propose a novel deep multi-view clustering framework to enhance its representation learning capabilities.This research did not involve human subjects, human-related data (e.g., personal identifiers, behavioral records), or animal subjects. This research did not receive any external sponsorship or funding, and none of the authors have any financial, professional, or personal conflicts of interest. Throughout this research, we strictly adhered to the principles of research integrity. All experimental procedures, data analysis, and result interpretation were performed in an objective and transparent manner, with full records maintained for verification. No ethical violations, such as data fabrication, manipulation, or plagiarism, occurred at any stage.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we have uploaded the source code. All datasets used in our experiments are from public datasets. In addition, all experimental procedures and result reports follow transparent standards. Section 4.1 of the main paper details the evaluation metrics (e.g., NMI, ACC, ARI) and includes the hyperparameter search range. To account for the randomness of model initialization and data partitioning, we report the average results of 10 independent runs.

REFERENCES

Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16752–16761, 2023.

Jinrong Cui, Yuting Li, Yulu Fu, and Jie Wen. Multi-view self-expressive subspace clustering network. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 417–425, 2023.

Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 12350–12368, 2023.

Hongchang Gao, Feiping Nie, Xuelong Li, and Heng Huang. Multi-view subspace clustering. In *Proceedings of the IEEE international conference on computer vision*, pp. 4238–4246, 2015.

Yuan Gao, Qian Zhao, Laurence T Yang, Jing Yang, and Jieming Yang. Tensor representation based multi-view graph contrastive learning for ioe intelligence. *IEEE Internet of Things Journal*, 2024.

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.

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

Shizhe Hu, Guoliang Zou, Chaoyang Zhang, Zhengzheng Lou, Ruilin Geng, and Yangdong Ye. Joint contrastive triple-learning for deep multi-view clustering. *Information Processing & Management*, 60(3):103284, 2023.

Weitian Huang, Sirui Yang, and Hongmin Cai. Generalized information-theoretic multi-view clustering. *Advances in neural information processing systems*, 36:58752–58764, 2023.

Guanzhou Ke, Zhiyong Hong, Zhiqiang Zeng, Zeyi Liu, Yangjie Sun, and Yannan Xie. Conan: contrastive fusion networks for multi-view clustering. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 653–660. IEEE, 2021.

Wei Lan, Tianchuan Yang, Qingfeng Chen, Shichao Zhang, Yi Dong, Huiyu Zhou, and Yi Pan. Multiview subspace clustering via low-rank symmetric affinity graph. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Lusi Li, Zhiqiang Wan, and Haibo He. Incomplete multi-view clustering with joint partition and graph learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):589–602, 2021.

Qilin Li, Senjian An, Ling Li, Wanquan Liu, and Yanda Shao. Multi-view diffusion process for spectral clustering and image retrieval. *IEEE Transactions on Image Processing*, 32:4610–4620, 2023.

Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2023.

Honglin Liu, Peng Hu, Changqing Zhang, Yunfan Li, and Xi Peng. Interactive deep clustering via value mining. *Advances in Neural Information Processing Systems*, 37:42369–42387, 2024.

Xinwang Liu, Xinzhong Zhu, Miaomiao Li, Lei Wang, Chang Tang, Jianping Yin, Dinggang Shen, Huaimin Wang, and Wen Gao. Late fusion incomplete multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2410–2423, 2018.

Xinwang Liu, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu. Efficient and effective regularized incomplete multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2634–2646, 2020.

Yue Liu, Wenxuan Tu, Sihang Zhou, Xinwang Liu, Linxuan Song, Xihong Yang, and En Zhu. Deep graph clustering via dual correlation reduction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 7603–7611, 2022.

Yue Liu, Xihong Yang, Sihang Zhou, Xinwang Liu, Siwei Wang, Ke Liang, Wenxuan Tu, and Liang Li. Simple contrastive graph clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 14193–14201, 2024.

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

Shirui Luo, Changqing Zhang, Wei Zhang, and Xiaochun Cao. Consistent and specific multi-view subspace clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Yazhou Ren, Jingyu Pu, Chenhang Cui, Yan Zheng, Xinyue Chen, Xiaorong Pu, and Lifang He. Dynamic weighted graph fusion for deep multi-view clustering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 4842–4850, 2024.

John Smith, Wenxuan Tu, Junlong Wu, Wenxin Zhang, Jingxin Liu, Haotian Wang, Jieren Cheng, Huajie Lei, Guangzhen Yao, Lingren Wang, et al. Dual boost-driven graph-level clustering network. *arXiv preprint arXiv:2504.05670*, 2025.

Huayi Tang and Yong Liu. Deep safe multi-view clustering: Reducing the risk of clustering performance degradation caused by view increase. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 202–211, 2022.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23976–23985, 2023.

Jing Wang, Songhe Feng, Gengyu Lyu, and Zhibin Gu. Triple-granularity contrastive learning for deep multi-view subspace clustering. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 2994–3002, 2023.

Zichen Wen, Yawen Ling, Yazhou Ren, Tianyi Wu, Jianpeng Chen, Xiaorong Pu, Zhifeng Hao, and Lifang He. Homophily-related: adaptive hybrid graph filter for multi-view graph clustering. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, pp. 15841–15849, 2024.

Wei Xia, Sen Wang, Ming Yang, Quanxue Gao, Jungong Han, and Xinbo Gao. Multi-view graph embedding clustering network: Joint self-supervision and block diagonal representation. *Neural Networks*, 145:1–9, 2022.

Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9234–9243, 2021.

Jie Xu, Chao Li, Yazhou Ren, Liang Peng, Yujie Mo, Xiaoshuang Shi, and Xiaofeng Zhu. Deep incomplete multi-view clustering via mining cluster complementarity. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8761–8769, 2022a.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16051–16060, 2022b.

Jie Xu, Shuo Chen, Yazhou Ren, Xiaoshuang Shi, Heng Tao Shen, Gang Niu, and Xiaofeng Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. In *Thirty-seventh Conference on Neural Information Processing Systems*, pp. 1119–1131, 2023.

Xiaoqiang Yan, Zhixiang Jin, Fengshou Han, and Yangdong Ye. Differentiable Information Bottleneck for Deterministic Multi-View Clustering . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27425–27434, 2024.

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

Chong You, Chun-Guang Li, Daniel P Robinson, and René Vidal. Oracle based active set algorithm for scalable elastic net subspace clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3928–3937, 2016.

Xuejiao Yu, Yi Jiang, Guoqing Chao, and Dianhui Chu. Deep contrastive multi-view subspace clustering with representation and cluster interactive learning. *IEEE Transactions on Knowledge and Data Engineering*, 37(1):188–199, 2025.

A Technical Appendices and Supplementary Material

A.1 Algorithm

The entire algorithm of MANGO is summarized in Algorithm 1

Algorithm 1 The algorithm of MANGO

Input: Multi-view data $\{\mathbf{X}^v \in \mathbb{R}^{n\times d_v}\}_{v=1}^m$; Training iterations E; Trade-off coefficients α; β and γ; diffusion steps T.
Output $\mathbf{A}_{final}$.
for $epoch = 1$ to E **do**
    Compute embeddings $\{\mathbf{Z}^v\}_{v=1}^m$ via $\{MLP^v\}_{v=1}^m$.
    Compute reconstruction loss$\mathcal{L}_{rec}$ through Eq. equation 1.
    Compute regularization loss$\mathcal{L}_{reg}$ through Eq. equation 2.
    Compute contrastive loss$\mathcal{L}_{inter}$ through Eq. equation 6.
    Compute consistency loss$\mathcal{L}_{consist}$ through Eq. equation 9.
    Update the network by optimizing $\mathcal{L}$ in Eq. equation 13.
end for
Build affinity matrix $\mathbf{A}$.
for $t = 1$ to T **do**
    Compute the information entropy of $\tilde{\mathbf{A}}^t$ through Eq. equation 10.
end for
Compute final affinity matrix$\mathbf{A}_{final}$ through Eq. equation 12
Performing the spectral clustering on $\mathbf{A}_{final}$ to obtain the final clustering results.

A.2 Additional experimental results

In this section, we provide complete results on all datasets, including parameter sensitivity analysis and ablation experiments. Table 4-Table 9 summarize the ablation studies of MANGO on all datasets for the three loss items and other improvements, while Figure 4 shows the sensitivity of MANGO to parameters α, β, and γ on all datasets.

As shown in Table 4 to Table 9, the complete MANGO model consistently outperforms its ablated variants across all datasets. This demonstrates that the integration of the self-expressive module, contrastive learning module, view consistency module, and adaptive diffusion module enables MANGO to fully exploit the rich information embedded in multi-view data, thereby enhancing clustering performance.

As for the parameter sensitivity analysis, Figure 4 demonstrates that MANGO consistently achieves stable and accurate clustering results across all 12 datasets over a broad range of parameter values, highlighting its robustness and practical reliability.

Table 4: Ablation study on Yale and ORL dataset

	$\mathcal{L}_{rec}$	$\mathcal{L}_{contra}$	$\mathcal{L}_{consist}$	random	diffusion	Yale			ORL		
						ACC	NMI	ARI	ACC	NMI	ARI
(a)	✓					0.260	0.306	0.058	0.923	0.963	0.892
(b)	✓	✓				0.670	0.708	0.517	0.930	0.960	0.893
(c)	✓	✓		✓		0.684	0.704	0.514	0.925	0.961	0.895
(d)	✓	✓		✓	✓	0.712	0.743	0.568	0.930	0.959	0.885
(e)	✓		✓			0.667	0.717	0.533	0.925	0.964	0.896
(f)	✓		✓		✓	0.694	0.744	0.567	0.933	0.964	0.904
(g)	✓	✓	✓	✓		0.687	0.716	0.526	0.931	0.964	0.897
(h)	✓	✓	✓	✓	✓	**0.716**	**0.751**	**0.584**	**0.943**	**0.966**	**0.911**

Table 5: Ablation study on BBC-Sport and Scene-15 dataset

	$\mathcal{L}_{rec}$	$\mathcal{L}_{contra}$	$\mathcal{L}_{consist}$	random	diffusion	BBC-Sport			Scene-15		
						ACC	NMI	ARI	ACC	NMI	ARI
(a)	✓					0.778	0.786	0.682	0.481	0.493	0.312
(b)	✓	✓				0.695	0.776	0.648	0.492	0.495	0.331
(c)	✓	✓		✓		0.701	0.764	0.590	0.493	0.498	0.327
(d)	✓	✓		✓	✓	0.959	0.871	0.912	0.493	0.494	0.333
(e)	✓		✓			0.432	0.200	0.107	0.477	0.493	0.321
(f)	✓		✓		✓	0.416	0.179	0.099	0.493	0.503	0.330
(g)	✓	✓	✓	✓		0.681	0.712	0.558	0.490	0.496	0.330
(h)	✓	✓	✓	✓	✓	**0.965**	**0.885**	**0.912**	**0.498**	**0.504**	**0.339**

Table 6: Ablation study on MSRC-v1 and Reuters dataset

	$\mathcal{L}_{rec}$	$\mathcal{L}_{contra}$	$\mathcal{L}_{consist}$	random	diffusion	MSRC-v1			Reuters		
						ACC	NMI	ARI	ACC	NMI	ARI
(a)	✓					0.770	0.758	0.662	0.502	0.347	0.267
(b)	✓	✓				0.795	0.780	0.699	0.510	0.347	0.269
(c)	✓	✓		✓		0.800	0.781	0.700	0.535	0.364	0.286
(d)	✓	✓		✓	✓	0.893	0.845	0.902	0.507	0.344	0.258
(e)	✓		✓			0.781	0.749	0.660	0.542	0.360	0.279
(f)	✓		✓		✓	0.790	0.770	0.681	0.546	0.360	0.270
(g)	✓	✓	✓	✓		0.863	0.818	0.748	0.551	0.340	0.261
(h)	✓	✓	✓	✓	✓	**0.950**	**0.908**	**0.887**	**0.587**	**0.376**	**0.287**

Table 7: Ablation study on LandUse-21 and Caltech101-20 dataset

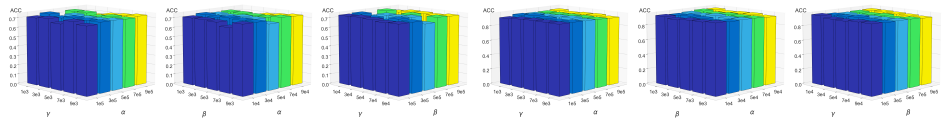
	$\mathcal{L}_{rec}$	$\mathcal{L}_{contra}$	$\mathcal{L}_{consist}$	random	diffusion	LandUse-21			Caltech101-20		
						ACC	NMI	ARI	ACC	NMI	ARI
(a)	✓					0.287	0.329	0.166	0.586	0.719	0.494
(b)	✓	✓				0.302	0.352	0.151	0.607	0.713	0.527
(c)	✓	✓		✓		0.317	0.349	0.160	0.632	0.738	0.544
(d)	✓	✓		✓	✓	0.289	0.330	0.136	0.639	0.724	0.552
(e)	✓		✓			0.324	0.345	0.160	0.552	0.687	0.460
(f)	✓		✓		✓	0.316	0.361	0.157	0.587	0.715	0.477
(g)	✓	✓	✓	✓		0.292	0.336	0.139	0.628	0.722	0.552
(h)	✓	✓	✓	✓	✓	**0.327**	**0.352**	**0.163**	**0.641**	**0.733**	**0.554**

Table 8: Ablation study on ALOI-100 and STL10 dataset

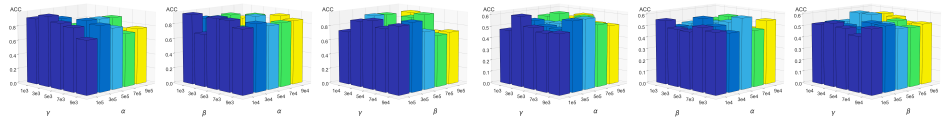
	$\mathcal{L}_{rec}$	$\mathcal{L}_{contra}$	$\mathcal{L}_{consist}$	random	diffusion	ALOI-100			STL10		
						ACC	NMI	ARI	ACC	NMI	ARI
(a)	✓					0.827	0.893	0.745	0.901	0.832	0.776
(b)	✓	✓				0.856	0.894	0.759	0.550	0.578	0.391
(c)	✓	✓		✓		0.872	0.906	0.789	0.841	0.815	0.759
(d)	✓	✓		✓	✓	0.868	0.902	0.776	0.866	0.820	0.708
(e)	✓		✓			0.868	0.901	0.774	0.702	0.753	0.548
(f)	✓		✓		✓	0.869	0.900	0.776	0.618	0.582	0.417
(g)	✓	✓	✓	✓		0.871	0.901	0.797	0.781	0.795	0.666
(h)	✓	✓	✓	✓	✓	**0.891**	**0.911**	**0.805**	**0.968**	**0.920**	**0.930**

Table 9: Ablation study on HandWritten and MNIST-3V dataset

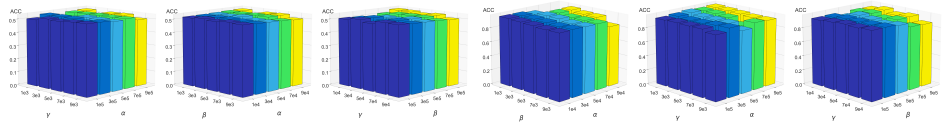
	$\mathcal{L}_{rec}$	$\mathcal{L}_{contra}$	$\mathcal{L}_{consist}$	random	diffusion	HandWritten			MNIST-3V		
						ACC	NMI	ARI	ACC	NMI	ARI
(a)	✓					0.967	0.926	0.926	0.949	0.928	0.898
(b)	✓	✓				0.971	0.935	0.935	0.988	0.966	0.974
(c)	✓	✓		✓		0.974	0.940	0.942	0.986	0.961	0.970
(d)	✓	✓		✓	✓	0.977	0.946	0.948	0.988	0.964	0.973
(e)	✓		✓			0.970	0.933	0.933	0.956	0.932	0.909
(f)	✓		✓		✓	0.977	0.944	0.948	0.988	0.964	0.973
(g)	✓	✓	✓	✓		0.970	0.934	0.934	0.988	0.965	0.973
(h)	✓	✓	✓	✓	✓	**0.978**	**0.948**	**0.951**	**0.989**	**0.966**	**0.975**



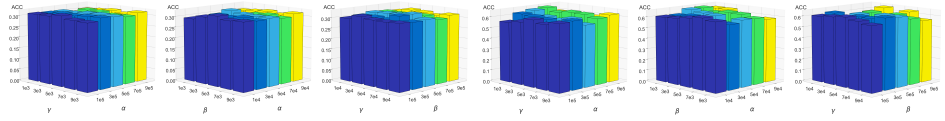
(a) Yale (b) ORL



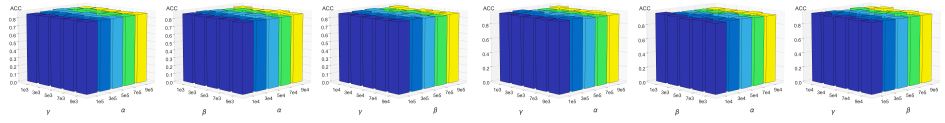
(c) BBC-Sport (d) Reuters



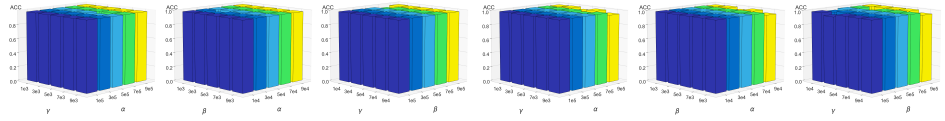
(e) Scene-15 (f) MSRC-v1



(g) LandUse-21 (h) Caltech101-20



(i) ALOI-100 (j) STL10



(k) HandWritten (l) MNIST-3V

Figure 4: Sensitivity Analysis of the MANGO Model to Parameters α, β and γ on Twelve Datasets

A.3 THE USE OF LARGE LANGUAGE MODELS

During this research and the writing of this paper, we incorporated a Large Language Model (LLM) as an auxiliary tool to improve text processing efficiency and facilitate preliminary literature search preparation. It should be clarified that this tool's use was strictly limited to auxiliary tasks and did not participate in the core aspects of this research, including but not limited to research design, primary data collection, experimental workflow, statistical analysis, and the derivation and demonstration of scientific conclusions. The scientific integrity, rigor, and originality of the core research content are the sole responsibility of the authors.

Specifically, the LLM's auxiliary role in this research focused on the following two aspects:

(1)Text Polishing and Grammar Standardization: Optimizing the language expression of selected paragraphs in the first draft of the paper primarily involved correcting grammatical errors, improving sentence fluency, and assisting with standardizing academic terminology to ensure that the text adheres to the language logic and formatting requirements of academic writing. The final text's academic content, logical structure, and core ideas were all reviewed and confirmed by the authors.

(2)Preliminary Literature Review Assistance: During the literature search phase, LLM assisted in generating a preliminary conceptual framework and keyword list for a specific research field, providing reference for the authors to determine the scope of their literature search and select their search strategy. It should be emphasized that all the literature included in the literature review of this study were read in full by the authors one by one, and their research relevance, content accuracy and academic value were independently verified before final determination. The literature recommendation results generated by the model were not directly used.