Mamba-Driven Multi-View Discriminative Clustering via Global-Local Cross-View Sequence Modeling

Yuanyang Zhang[1], Xinhang Wan[2], Chao Zhang[3], Jie Xu[4], Cunjian Chen[5], Tien-Tsin Wong[5], Li Yao[1,6*], Yijie Lin[7*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]College of Systems Engineering, National University of Defense Technology, Changsha, China
[3]School of Robotics and Automation, Nanjing University, Suzhou, China
[4]Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore
[5]Department of Data Science & AI, Faculty of Information Technology, Monash University, Melbourne, Australia
[6]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[7]The Hong Kong University of Science and Technology, Hong Kong SAR, China
zhangyuanyang@seu.edu.cn, yao.li@seu.edu.cn, linyijie.gm@gmail.com

Abstract

Multi-view clustering (MVC) has recently garnered increasing attention for its ability to partition unlabeled samples into distinct clusters by leveraging complementary and consistent information from different views. Existing MVC methods primarily combine deep neural networks with contrastive learning for cross-view representation learning, yet often overlook the inherent global-local structural relationships among samples. While GNN-based methods capture local structures, they struggle to model global dependencies, leading to inferior inter-cluster separability. In contrast, Transformer-based methods excel at global aggregation but suffer from quadratic complexity, and their attention smoothing effect weakens fine-grained local structures, resulting in suboptimal intra-cluster compactness. To address these limitations, we propose a novel end-to-end MVC framework called Mamba-Driven Multi-View Discriminative Clustering via Global-Local Cross-View Sequence Modeling (MGLC). By flexibly constructing multi-view sequences, MGLC fully exploits the efficient sequence modeling capabilities of Mamba to jointly model cross-view dependencies and global-local structural relationships among samples. Furthermore, MGLC introduces a Cross-Mamba Fusion module to dynamically integrate cross-view and global-local structural representations. Additionally, MGLC incorporates a Dual Calibration Contrastive Learning module, guided by high-confidence pseudo-labels, that adaptively refines both feature and semantic representations while mitigating false negatives among semantically similar samples. Extensive comparative experiments and ablation studies demonstrate the effectiveness of MGLC.

Introduction

Multi-view data captures the complementary and consistent information of the same object from different sources or modalities (Lu et al. 2024), and has been widely applied in practical scenarios such as medical analysis (Kim and Park 2024; Li et al. 2025) and autonomous driving (Xing et al. 2025). As one of the most effective tools for analyzing multi-view data, multi-view clustering (MVC) partitions unlabeled samples into distinct semantic clusters by mining and leveraging the latent information across different views (Lin et al. 2022, 2021; Yang et al. 2022; Zhu et al. 2025b; Zhang et al. 2025c; Wan et al. 2024; Guan et al. 2025b,a).

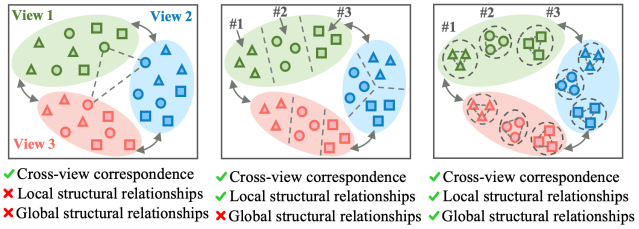

Figure 1: Our key idea. Achieving inter-cluster separability and intra-cluster compactness depends on effectively leveraging three essential types of representational information: cross-view correspondence, global structural relationships, and local structural relationships. However, most existing MVC methods overlook the intrinsic global-local structural relationships, resulting in suboptimal clustering performance.

The core objective of MVC is to achieve inter-cluster separability and intra-cluster compactness (Zhang et al. 2025b; Sun et al. 2025; Zhang et al. 2025a; Li et al. 2023; Xu et al. 2025; Jiang et al. 2025). This objective hinges on the effective exploitation of three key types of representational information. As illustrated in Figure 1, cross-view correspondence captures the semantic consistency of the same instance across different views; global structural relationships characterize the macro-level distribution of samples, enabling clearer separation between clusters; and local structural relationships preserve fine-grained neighborhood similarities, promoting intra-cluster coherence. The synergy of these components facilitates high-quality clustering.

The predominant strategy for achieving inter-cluster sep-

*Corresponding authors: Li Yao; Yijie Lin.

arability and intra-cluster compactness is to combine deep neural networks and contrastive learning. Researchers have explored several deep MVC approaches, including autoencoder (AE)-based methods (Xu et al. 2022), graph neural network (GNN)-based methods (Chao et al. 2025), and Transformer-based methods (Yan et al. 2023; Wang et al. 2025). Despite significant progress, these methods often overlook the inherent global-local structural relationships among samples. Specifically, AE-based methods utilize contrastive learning for cross-view alignment but ignore both global and local structures, limiting clustering discriminability. While GNN-based methods effectively capture local neighborhood structures, they struggle to model global dependencies, resulting in inferior inter-cluster separability. In contrast, Transformer-based models excel at global aggregation but suffer from quadratic complexity, and their attention smoothing effect weakens fine-grained local structures, leading to suboptimal intra-cluster compactness.

The efficient state-space modeling framework Mamba (Gu and Dao 2023) offers input adaptivity and global modeling capabilities through its selective mechanism, enabling effective capture of critical information. Meanwhile, its linear computational complexity significantly reduces computational overhead and greatly enhances inference efficiency, making it a promising alternative. For example, Zhu et al. (Zhu et al. 2025a) employ Mamba for trustworthy cross-view fusion. However, it still focuses solely on cross-view learning and has not systematically utilized Mamba's powerful sequence modeling potential to integrate both global and local structures.

To address these challenges, we propose a novel MVC framework called Mamba-Driven Multi-View Discriminative Clustering via Global-Local Cross-View Sequence Modeling (MGLC). As shown in Figure 2, our approach first constructs flexible multi-view sequences to fully leverage Mamba's efficient sequence modeling capabilities for jointly modeling cross-view dependencies and global-local structural relationships among samples. The proposed Cross-Mamba Fusion module then dynamically fuses cross-view and global-local representations to enhance feature discriminability. Moreover, we introduce a pseudo-label-guided Dual Calibration Contrastive Learning module that adaptively refines both feature and semantic representations, while mitigating false negatives among semantically similar samples. Our contributions are summarized as follows:

- We propose a novel end-to-end MVC framework, MGLC, which fully leverages Mamba's efficient sequence modeling capabilities to jointly model cross-view dependencies and global-local structural relationships through the flexible construction of multi-view sequences.
- We design a Cross-Mamba Fusion module to dynamically integrate cross-view and global-local structural representations. In addition, we develop a Dual Calibration Contrastive Learning module that adaptively refines both feature and semantic representations while effectively reducing false negatives among semantically similar samples.
- Extensive experiments on eight benchmark datasets demonstrate that MGLC consistently outperforms state-of-the-art MVC methods in both clustering performance and computational efficiency.

Related Work

Deep Multi-view Clustering

Inspired by the powerful feature representation capabilities of deep neural networks, many deep MVC methods have been developed. These methods are primarily categorized as follows: (a) AE-based methods (Xu et al. 2021; Yan et al. 2024) leverage contrastive learning based on autoencoder architectures for cross-view alignment, but often overlook global and local structures, thereby hindering clustering discriminability; (b) GNN-based methods (Ren et al. 2024; Chao et al. 2025) construct intra-view or cross-view graphs and aggregate features via message passing to model local structural dependencies. However, their inherently local propagation limits the ability to capture long-range global dependencies, resulting in inferior inter-cluster separability; (c) Transformer-based methods (Yan et al. 2023; Wang et al. 2025) enable global feature aggregation via self-attention, but their quadratic complexity limits efficiency, especially in large-scale scenarios. In addition, attention smoothing weakens fine-grained local structures, leading to insufficient intra-cluster compactness.

State Space Model

The state space model (SSM) with linear complexity offers a promising approach for modeling long-range dependencies. Moreover, the Structured State Space Sequence (S4) model (Gu, Goel, and Ré 2021) improves computational efficiency through a novel parameterization while preserving its theoretical advantages. Building upon this foundation, Mamba (Gu and Dao 2023) stands out by introducing a data-dependent SSM layer and a parallel scan (S6) selection mechanism, surpassing Transformers (Vaswani et al. 2017) in both inference speed and overall performance. Zhu et al. (Zhu et al. 2025a) leverage Mamba for trustworthy cross-view fusion, but their method focuses solely on cross-view learning and overlooks Mamba's potential in modeling global and local structures among samples.

The Proposed Method

Notations. Given a multi-view dataset $\{\boldsymbol{X}^v \in \mathbb{R}^{N\times D_v}\}_{v=1}^{V}$ with N samples across V views, where $\boldsymbol{X}^v$ denotes the data matrix of the v-th view and D_v is the feature dimensionality. Multi-view clustering aims to partition the N samples into C clusters.

Preliminaries

The SSM-based models Mamba (Gu and Dao 2023) and Vision Mamba (Vim) (Zhu et al. 2024) are inspired by continuous systems, which map a one-dimensional sequence $x(t) \in \mathbb{R} \mapsto y(t) \in \mathbb{R}$ through an N-dimensional hidden state $h(t) \in \mathbb{R}^N$. The hidden state evolves over time with parameters $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, following linear ordinary differential equations:

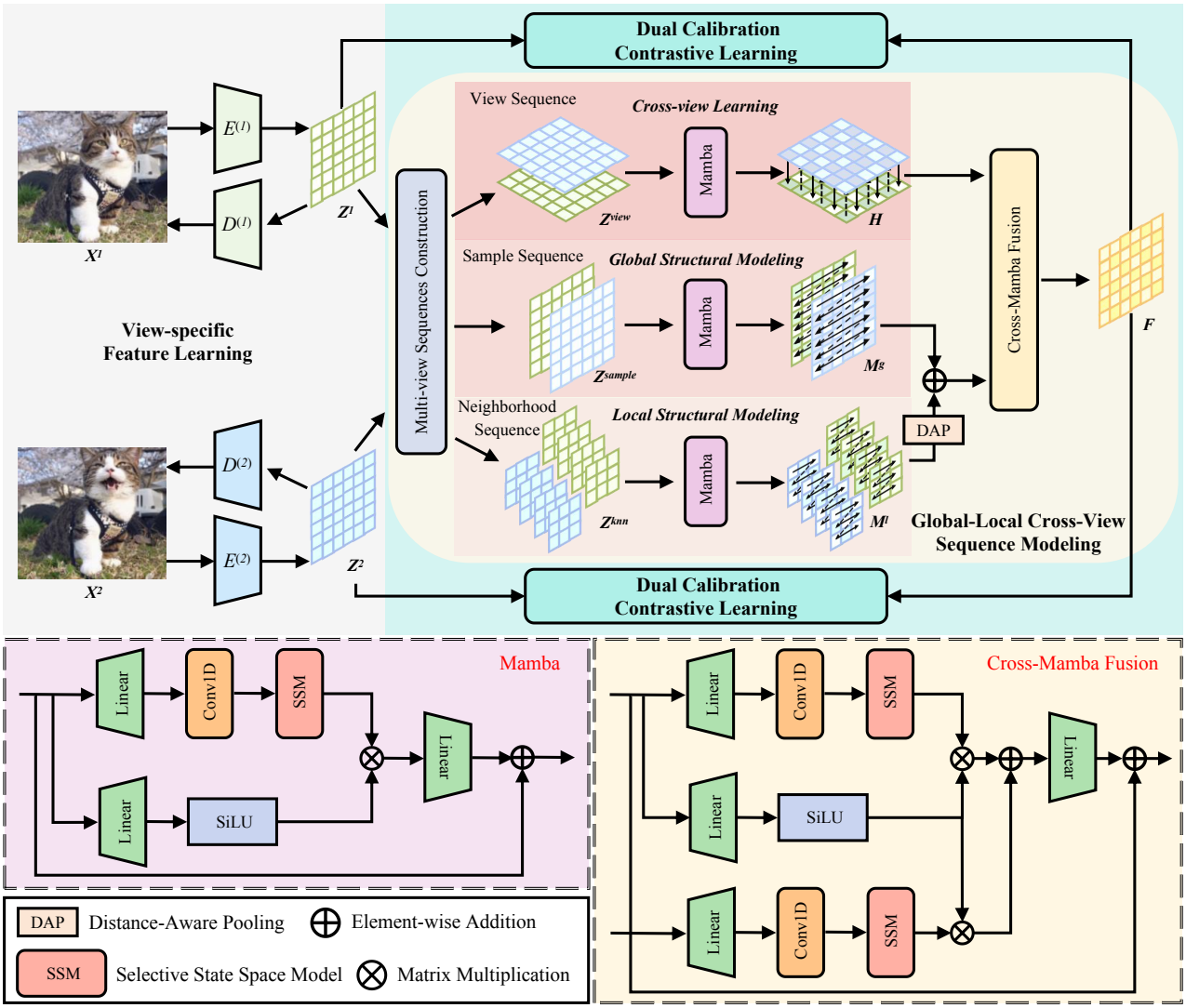

Figure 2: Without loss of generality, we take bi-view data as a showcase to demonstrate the overall framework of our proposed MGLC. As shown, our method is mainly divided into three modules: (1) View-specific Feature Learning; (2) Global-Local Cross-View Sequence Modeling; (3) Dual Calibration Contrastive Learning. Note that, *E*: Encoder; *D*: Decoder.

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the state matrix, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, and $\mathbf{C} \in \mathbb{R}^{1\times N}$ are projection parameters.

To adapt SSM for deep learning, the system is discretized using zero-order hold (Pechlivanidou and Karampetakis 2022). The continuous parameters $\mathbf{A}$, $\mathbf{B}$ are transformed into their discrete counterparts $\overline{\mathbf{A}} \in \mathbb{R}^{N\times N}$, $\overline{\mathbf{B}} \in \mathbb{R}^{N\times 1}$ using a timescale parameter $\Delta \in \mathbb{R}$:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta\mathbf{A}), \\ \overline{\mathbf{B}} &= (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B} \approx \Delta\mathbf{B}. \end{aligned} \tag{2}$$

Thus, the discrete SSM can be expressed as:

$$\begin{aligned} h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \\ y_t &= \mathbf{C}h_t, \end{aligned} \tag{3}$$

where $h_t, h_{t-1} \in \mathbb{R}^{N\times 1}$ and $x_t \in \mathbb{R}$.

View-specific Feature Learning

Deep autoencoders are commonly used for unsupervised representation learning (Xu et al. 2022; Sun et al. 2024) by minimizing the reconstruction error. Given that different views inherently contain view-specific information, we employ a dedicated autoencoder for each view to learn its latent representation $\boldsymbol{Z}^v \in \mathbb{R}^{N\times d}$ by minimizing the reconstruction loss:

$$\mathcal{L}_{\text{REC}} = \sum_{v=1}^{V}\sum_{i=1}^{N}\left\|\boldsymbol{X}_i^v - D_{\phi^v}^{(v)}\left(\boldsymbol{Z}_i^v\right)\right\|_2^2, \tag{4}$$

where $\boldsymbol{X}_i^v$ is the i-th sample from the v-th view, and $D_{\phi^v}^{(v)}$ denotes the corresponding decoder with parameters ϕ^v. The latent representation $\boldsymbol{Z}_i^v$ is obtained as follows:

$$\boldsymbol{Z}_i^v = E_{\eta^v}^{(v)}\left(\boldsymbol{X}_i^v\right), \tag{5}$$

where $E_{\eta^v}^{(v)}$ is the encoder of v-th view with parameters η^v.

Global-Local Cross-View Sequence Modeling

To fully exploit cross-view information and the global-local structural relationships among samples for discriminative clustering, we design a Global-Local Cross-View Sequence Modeling (GLCSM) module, inspired by the Mamba architecture (Gu and Dao 2023). Specifically, the module constructs flexible multi-view sequences including view, sample, and neighborhood sequences to leverage Mamba's selective state-space mechanism for jointly modeling cross-view dependencies and global-local structures. The proposed Cross-Mamba Fusion module then dynamically integrates cross-view and global-local structural representations. The GLCSM module comprises Cross-view Learning, Global Structural Modeling, Local Structural Modeling, and Cross-Mamba Fusion.

Cross-view Learning. We concatenate the representations from all views to form a view sequence:

$$\boldsymbol{Z}^{\text{view}} = \left[\boldsymbol{Z}^1, \boldsymbol{Z}^2, \ldots, \boldsymbol{Z}^V\right] \in \mathbb{R}^{N\times V\times d}, \tag{6}$$

which is fed into the Mamba block for adaptive cross-view aggregation via the Selective State Space Model (SSM). The Mamba block consists of two parallel branches: the upper branch applies a linear projection followed by a 1D convolution and SSM; the lower branch applies a linear transformation and SiLU activation for non-linear mapping. The two outputs are fused through element-wise multiplication, followed by a linear projection and residual connection:

$$\begin{aligned}
\boldsymbol{H}_t &= \text{SiLU}\left(\text{Linear}(\boldsymbol{Z}^{\text{view}})\right), \\
\boldsymbol{H}_b &= \text{SSM}\left(\text{Conv1D}(\text{Linear}(\boldsymbol{Z}^{\text{view}}))\right), \\
\boldsymbol{H} &= \text{Linear}(\boldsymbol{H}_t \otimes \boldsymbol{H}_b) + \boldsymbol{Z}^{\text{view}}, \boldsymbol{H} \in \mathbb{R}^{N\times V\times d}.
\end{aligned} \tag{7}$$

where $\boldsymbol{H}$ denotes the cross-view representation.

Global Structural Modeling. To capture global structural relationships among samples, we construct a sample sequence by concatenating the representations of all N samples in each view:

$$\boldsymbol{Z}^{\text{sample}} = [\boldsymbol{z}_1^v, \boldsymbol{z}_2^v, \ldots, \boldsymbol{z}_N^v]_{v=1}^V \in \mathbb{R}^{V\times N\times d}, \tag{8}$$

where $\boldsymbol{z}_i^v$ denotes the feature representation of the i-th sample in the v-th view. This sequence is then passed through the Mamba block to model long-range dependencies (Lin et al. 2024) across samples:

$$\boldsymbol{M}^g = \text{Mamba}(\boldsymbol{Z}^{\text{sample}}) \in \mathbb{R}^{V\times N\times d}, \tag{9}$$

where $\boldsymbol{M}^g$ denotes the global structural representation.

Local Structural Modeling. To capture fine-grained local structural relationships, we construct a K-nearest neighbor graph for each view. For each sample i, we identify its K nearest neighbors $\mathcal{N}_i^{(k)}$ and form the local neighborhood sequence:

$$\boldsymbol{Z}_i^{\text{knn}} = \left\{\boldsymbol{z}_j \mid j \in \mathcal{N}_i^{(k)}\right\}, \quad \boldsymbol{Z}^{\text{knn}} \in \mathbb{R}^{V\times N\times K\times d}. \tag{10}$$

The neighborhood sequence is then passed through the Mamba block to model structural dependencies within each local region:

$$\boldsymbol{M}^l = \text{Mamba}\left(\boldsymbol{Z}^{\text{knn}}\right) \in \mathbb{R}^{V\times N\times K\times d}, \tag{11}$$

where N, K, and d denote the number of samples, neighborhood size, and feature dimension, respectively.

To emphasize more relevant local information, we apply a distance-aware pooling strategy. The weight for each neighbor is computed based on Euclidean distance:

$$\boldsymbol{w}_{ij} = \frac{\exp(-\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2)}{\sum_{j'\in\mathcal{N}_i}\exp(-\|\boldsymbol{z}_i - \boldsymbol{z}_{j'}\|_2)}. \tag{12}$$

The aggregated local structural representation for sample i is then obtained by weighted summation:

$$\hat{\boldsymbol{M}}_i^l = \sum_{j\in\mathcal{N}_i} \boldsymbol{w}_{ij}\cdot\boldsymbol{M}_{i,j}^l, \quad \hat{\boldsymbol{M}}^l \in \mathbb{R}^{V\times N\times d}, \tag{13}$$

Finally, the global $\boldsymbol{M}^g$ and local $\hat{\boldsymbol{M}}^l$ structural representations are fused by element-wise addition to form the final global-local representation $\boldsymbol{M} = \boldsymbol{M}^g + \hat{\boldsymbol{M}}^l$. Then $\boldsymbol{M}$ is fed into the Cross-Mamba Fusion module.

Cross-Mamba Fusion. To dynamically fuse cross-view and global-local structural representations for generating discriminative cluster embeddings, we design a Cross-Mamba Fusion module. Specifically, the cross-view representation $\boldsymbol{H}$ and the sample-level global-local structural representation $\boldsymbol{M}$ are first processed by parallel linear projection, 1D convolution, SiLU activation, and Selective SSM to obtain state outputs $\boldsymbol{F}_H$ and $\boldsymbol{F}_M$. We then project $\boldsymbol{H}$ via a linear layer to obtain a shared feature representation:

$$\boldsymbol{e} = \text{Linear}(\boldsymbol{H}), \tag{14}$$

This shared representation is activated by a SiLU function to produce dynamic weights, which modulate the two state outputs in an element-wise manner, enhancing the selectivity and stability of the fusion process:

$$\boldsymbol{F}'_H = \boldsymbol{F}_H \otimes \text{SiLU}(\boldsymbol{e}), \quad \boldsymbol{F}'_M = \boldsymbol{F}_M \otimes \text{SiLU}(\boldsymbol{e}). \tag{15}$$

Finally, the modulated outputs are fused via linear projection and residual connection to produce the final unified representation $\boldsymbol{F} \in \mathbb{R}^{N\times d}$:

$$\boldsymbol{F} = \text{Reshape}\left(\text{Linear}(\boldsymbol{F}'_H + \boldsymbol{F}'_M) + \boldsymbol{F}'_H\right). \tag{16}$$

Dual Calibration Contrastive Learning

To dynamically refine both feature-level and semantic-level representations, we design a pseudo-label-guided dual calibration contrastive loss. Unlike traditional contrastive learning, which may mistakenly treat semantically similar samples as negatives, our method leverages high-confidence

pseudo labels to mitigate such false negatives (Yang et al. 2021; Lin et al. 2023), thereby improving clustering performance.

Specifically, based on the fused representation $\boldsymbol{F}$ and view-specific features $\boldsymbol{Z}^v$ obtained via the encoder, we first obtain pseudo labels:

$$\boldsymbol{Q} = \mathrm{Classifier}(\boldsymbol{F}), \quad \boldsymbol{Q}^v = \mathrm{Classifier}(\boldsymbol{Z}^v), \tag{17}$$

where $\mathrm{Classifier}(\cdot)$ denotes the shared classification head.

Based on these outputs, we construct a high-confidence pseudo label graph $\boldsymbol{W} \in \mathbb{R}^{N \times N}$ as follows:

$$\boldsymbol{W}_{ij} = \begin{cases} 1 & \text{if } i = j, \\ \boldsymbol{Q}_i \cdot \boldsymbol{Q}_j^v & \text{if } i \neq j \text{ and } \boldsymbol{Q}_i \cdot \boldsymbol{Q}_j^v \geq \xi, \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

where ξ is a similarity threshold. Diagonal elements represent the same sample across different views. For off-diagonal elements, if the similarity is below ξ, the corresponding nodes are not connected in the pseudo label graph.

We then compute a cross-view feature similarity graph using cosine similarity:

$$\boldsymbol{S}_{ij} = \frac{\langle \boldsymbol{F}_i, \boldsymbol{Z}_j^v \rangle}{\|\boldsymbol{F}_i\|_2 \cdot \|\boldsymbol{Z}_j^v\|_2}, \tag{19}$$

where $\boldsymbol{S}_{ij}$ denotes the cosine similarity between the i-th fused feature $\boldsymbol{F}_i$ and the j-th view-specific feature $\boldsymbol{Z}_j^v$.

To encourage consistency between semantically similar yet distinct samples, we introduce a feature-level calibration contrastive loss that aligns the pseudo label graph $\boldsymbol{W}$ with the feature similarity graph $\boldsymbol{S}$:

$$\begin{aligned} \mathcal{L}_{\mathrm{FCC}} = &- \sum_{i=1}^{N} \boldsymbol{W}_{ii} \log \left(\frac{\exp(\boldsymbol{S}_{ii})}{\sum_j \exp(\boldsymbol{S}_{ij})} \right) \\ &- \sum_{i=1}^{N} \sum_{j \neq i} \boldsymbol{W}_{ij} \log \left(\frac{\exp(\boldsymbol{S}_{ij})}{\sum_j \exp(\boldsymbol{S}_{ij})} \right). \end{aligned} \tag{20}$$

The first term encourages self-alignment across views, while the second promotes the aggregation of semantically similar but distinct samples. During training, the high-confidence pseudo label graph $\boldsymbol{W}$ guides the learning of cross-view feature similarities.

To further enhance semantic consistency and enable robust end-to-end clustering, we introduce a category-level contrastive loss between the fused predictions $\boldsymbol{Q}$ and the view-specific predictions $\boldsymbol{Q}^v$:

$$\mathcal{L}_{\mathrm{ccl}} = -\frac{1}{2C} \sum_{i=1}^{C} \sum_{v=1}^{V} \log \frac{\exp(\mathrm{sim}(\boldsymbol{Q}_{\cdot i}, \boldsymbol{Q}_{\cdot i}^v)/\tau)}{\sum_{c=1}^{C} \exp(\mathrm{sim}(\boldsymbol{Q}_{\cdot i}, \boldsymbol{Q}_{\cdot c}^v)/\tau)}, \tag{21}$$

where τ is a temperature parameter (default set to 1) controlling the distribution sharpness, and $\mathrm{sim}(\cdot, \cdot)$ denotes the cosine similarity between class distributions.

To further improve the reliability of pseudo labels, we introduce a KL divergence-based self-supervised refinement mechanism. It encourages high-confidence predictions to not only improve their own representations but also guide the optimization of other samples through knowledge transfer. Specifically, for each instance, we compute the element-wise maximum between $\boldsymbol{Q}$ and $\boldsymbol{Q}^v$ to obtain the high-confidence pseudo labels:

$$\boldsymbol{q}_{ij} = \max\{\boldsymbol{Q}_{ij}, \boldsymbol{Q}_{ij}^v\}. \tag{22}$$

These are then normalized to construct a target distribution:

$$\boldsymbol{p}_{ij} = \frac{\boldsymbol{q}_{ij}^2}{\sum_{j=1}^{C} \boldsymbol{q}_{ij}^2}. \tag{23}$$

We optimize the clustering results by minimizing the KL divergence between the target distribution $\boldsymbol{P}$ and the current prediction $\boldsymbol{Q}$:

$$\mathcal{L}_{\mathrm{hg}} = \mathrm{KL}(\boldsymbol{P} \| \boldsymbol{Q}) = \sum_{i=1}^{N} \sum_{j=1}^{C} \boldsymbol{p}_{ij} \log \frac{\boldsymbol{p}_{ij}}{\boldsymbol{q}_{ij}}. \tag{24}$$

The final semantic-level calibration contrastive loss combines both objectives:

$$\mathcal{L}_{\mathrm{SCC}} = \mathcal{L}_{\mathrm{ccl}} + \mathcal{L}_{\mathrm{hg}}. \tag{25}$$

Objective Function

Our model is trained end-to-end and does not rely on post-processing with k-means clustering (Bauckhage 2015) to obtain the final cluster assignments. Upon convergence, the cluster label of the i-th sample ($1 \leq i \leq N$) is inferred as:

$$y_i = \arg \max_{1 \leq j \leq C} P_{ij}, \tag{26}$$

where $\boldsymbol{P}_{i,:} \in \mathbb{R}^C$ denotes the predicted assignment distribution of sample i. This enables direct label prediction without additional clustering steps.

The overall objective function is described as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{REC}} + \lambda_1 \mathcal{L}_{\mathrm{FCC}} + \lambda_2 \mathcal{L}_{\mathrm{SCC}}, \tag{27}$$

where λ_1 and λ_2 are trade-off coefficients.

Experiments

Experimental Setup

Datasets. We conducted experiments on eight benchmark datasets: BDGP (Cai et al. 2012), Mfeat[1], HW (Asuncion and Newman 2007), Aloideep (Geusebroek, Burghouts, and Smeulders 2005), NoisyMNIST (Wang et al. 2015), YouTubeFace[2], WebKB (Sun et al. 2007), and 100leaves[3]. Detailed descriptions of these datasets are provided in the supplementary materials. For a comprehensive analysis, we adopt three widely-used clustering metrics: Accuracy (ACC), Normalized Mutual Information (NMI), and Purity (PUR). Higher values on these metrics indicate better clustering performance.

Implementation Details. All experiments were conducted on a Linux system equipped with an Intel(R) Xeon(R) Silver

[1] https://archive.ics.uci.edu/dataset/72/multiple+features

[2] https://www.cs.tau.ac.il/~wolf/ytfaces/

[3] https://archive.ics.uci.edu/dataset/241/one+hundred+plant+species+leaves+data+set

Method	BDGP			Mfeat			HW			Aloideep		
	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
SiMVC	70.43	54.58	72.32	81.15	81.80	84.10	87.60	87.80	87.60	58.60	92.00	58.60
CoMVC	80.25	67.03	80.33	85.90	81.54	85.90	90.80	89.40	90.80	66.30	94.10	66.30
MFLVC	98.30	95.10	98.30	82.90	82.73	82.90	77.05	75.75	77.05	82.93	96.42	82.93
GCFagg	98.70	96.13	98.70	84.30	76.63	84.30	95.50	90.44	95.50	90.25	97.34	91.03
DealMVC	98.80	96.31	98.80	74.40	78.14	74.90	81.15	80.15	81.15	82.84	96.27	82.84
SEM	98.82	96.22	98.82	76.75	73.56	76.75	77.00	74.68	77.00	91.86	96.43	93.98
TMCN	74.12	50.42	74.12	81.85	80.42	81.85	86.21	83.11	85.45	89.35	95.96	90.84
GHICMC	85.52	69.57	85.52	94.86	91.72	94.86	97.44	94.16	97.44	90.28	97.02	90.28
MGLC	**99.12**	**97.33**	**99.12**	**97.26**	**93.75**	**97.26**	**97.83**	**94.66**	**97.83**	**95.22**	**98.30**	**95.26**
Δ	+0.30	+1.02	+0.30	+2.40	+2.03	+2.40	+0.39	+0.50	+0.39	+3.36	+0.96	+1.28

Table 1: Clustering results (%) across four multi-view benchmark datasets. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	NoisyMNIST			YouTubeFace			WebKB			100leaves		
	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
SiMVC	87.10	83.20	88.60	27.65	24.82	26.62	59.25	16.58	59.25	56.06	80.99	53.84
CoMVC	95.50	90.70	96.70	26.74	28.53	26.74	62.32	21.65	62.32	58.56	81.23	58.93
MFLVC	99.27	97.70	99.27	27.70	29.52	32.97	67.25	24.56	67.25	28.12	71.01	28.12
GCFagg	97.00	87.13	97.00	32.62	32.89	40.07	64.04	29.85	75.37	86.50	94.54	87.97
DealMVC	99.50	98.32	99.50	23.76	22.02	29.93	52.22	7.57	57.14	77.37	90.28	77.37
SEM	58.93	65.21	58.93	31.30	31.00	39.63	65.02	36.50	75.38	83.50	94.05	86.81
TMCN	96.68	93.37	97.97	30.39	31.63	38.86	61.58	21.57	67.98	66.50	83.39	68.63
GHICMC	97.68	87.35	97.68	31.43	31.53	38.85	71.43	30.10	71.43	93.85	96.82	93.85
MGLC	**99.69**	**98.95**	**99.69**	**34.52**	**34.88**	**43.23**	**76.85**	**48.84**	**81.77**	**95.21**	**97.25**	**95.21**
Δ	+0.19	+0.63	+0.19	+1.90	+1.99	+3.16	+5.42	+12.34	+6.39	+1.36	+0.43	+1.36

Table 2: Clustering results (%) across four multi-view benchmark datasets. The best and second-best results are highlighted in **bold** and underlined, respectively.

4215R CPU @ 3.20 GHz, 220 GB RAM, and an NVIDIA RTX 3090 GPU. For all datasets, the model was first warmed up using reconstruction loss for 200 epochs, followed by 100 epochs of training with the overall loss. We employed the Adam optimizer (Kingma and Ba 2014) with default settings in PyTorch (Paszke et al. 2019), using a learning rate of 0.0003 and a batch size of 256. For all comparison methods, we used the official implementations and ran them on our machine following the authors' recommended settings.

Comparison with State-of-the-arts

We compared MGLC with eight state-of-the-art MVC methods, including SiMVC (Trosten et al. 2021), CoMVC (Trosten et al. 2021), MFLVC (Xu et al. 2022), GCFagg (Yan et al. 2023), DealMVC (Yang et al. 2023), SEM (Xu et al. 2024), TMCN (Zhu et al. 2025a), and GHICMC (Chao et al. 2025). A detailed description of these baselines is provided in the supplementary materials.

Table 1 and Table 2 report the average clustering performance over five independent runs. The results reveal the following observations: (1) **Overall Superiority.** MGLC consistently outperforms all baseline methods across all datasets. For instance, on the WebKB dataset, it achieves improvements of 5.42%, 12.34%, and 6.39% in ACC, NMI, and PUR, respectively, over the second-best method, demonstrating the effectiveness of the proposed architecture in capturing cross-view and structural information. (2) **Scalability.** MGLC maintains strong performance on large-scale datasets. On the YouTubeFace dataset, it surpasses the second-best method by 1.90%, 1.99%, and 3.16% in ACC, NMI, and PUR, respectively, indicating superior scalability and robustness in large-scale scenarios.

Ablation Study

Ablation of GLCSM Module. To evaluate the effectiveness of the four key components within the proposed GLCSM module, including Cross-view Learning (CVL), Global Structural Modeling (GSM), Local Structural Modeling (LSM), and Cross-Mamba Fusion (CMF), we design four ablated variants: (1) *w/o* **CVL**: removes cross-view learning; (2) *w/o* **GSM**: excludes global structural modeling; (3) *w/o* **LSM**: drops local structural modeling; and (4) *w/o* **CMF**: replaces the Cross-Mamba Fusion with concatenation. As shown in Table 3, we can conclude that:

- Removing any single component results in noticeable performance drops, validating the necessity of each part in the GLCSM module.
- Among them, removing CVL or CMF results in the most significant performance drop, underscoring the impor-

Components	WebKB	HW	YouTubeFace
w/o **CVL**	70.82	92.29	32.60
w/o **GSM**	72.25	94.85	32.79
w/o **LSM**	74.88	95.65	33.63
w/o **CMF**	70.76	93.32	32.53
MGLC	**76.85**	**97.83**	**34.52**

Table 3: Ablation results on ACC (%) for Cross-view Learning (CVL), Global Structural Modeling (GSM), Local Structural Modeling (LSM), and Cross-Mamba Fusion (CMF).

Component			WebKB	HW	YouTubeFace
$\mathcal{L}_{\mathrm{FCC}}$	$\mathcal{L}_{\mathrm{SCC}}$	$\boldsymbol{W}$			
✗	✓	✗	73.12	96.95	32.58
✓	✗	✗	70.98	94.62	31.35
✓	✗	✓	72.28	95.19	31.86
✓	✓	✓	**76.85**	**97.83**	**34.52**

Table 4: Ablation results on ACC (%) for feature-level calibration contrastive loss ($\mathcal{L}_{\mathrm{FCC}}$), semantic-level calibration contrastive loss ($\mathcal{L}_{\mathrm{SCC}}$), and high-confidence pseudo-label graph ($\boldsymbol{W}$). ✗ and ✓ denote without or with the component.

tance of adaptive cross-view learning and dynamic fusion.

- LSM and GSM also contribute positively, confirming that modeling both global and local sample structures is crucial for learning discriminative cluster assignments.

Influence of Loss Components. As shown in Table 4, removing any loss function leads to a decline in clustering performance. In addition, removing the high-confidence pseudo label graph $\boldsymbol{W}$ from $\mathcal{L}_{\mathrm{FCC}}$ also degrades the clustering performance. These results indicate that $\mathcal{L}_{\mathrm{FCC}}$ achieves interaction alignment at the feature level, while $\mathcal{L}_{\mathrm{SCC}}$ enables interaction alignment at the semantic level. Moreover, the high-confidence pseudo label graph $\boldsymbol{W}$ helps reduce false negative samples among semantically similar samples.

Hyper-Parameters Analysis

We experimentally evaluated the impact of hyperparameters λ_1, λ_2, K, and ξ. As shown in Fig. 3, MGLC demonstrates robustness to variations in λ within the range $[10^{-1}, 10^{1}]$. Fig. 4 presents the analyses of K and ξ, where the optimal performance is achieved at $K = 8$, and the model benefits from increasing ξ, demonstrating the importance of high-confidence pseudo label graphs in reducing false negatives. Based on these results, we set $\lambda_1 = \lambda_2 = 1$, $K = 8$, and $\xi = 0.8$ in our experiments.

Runtime Analysis

To validate the efficiency of MGLC, we conducted a runtime analysis. Table 5 summarizes runtime and performance on the large-scale YouTubeFace dataset. Compared with existing baselines, MGLC achieves a better balance between efficiency and effectiveness, benefiting from Mamba's input-adaptive selection mechanism, which dynamically adjusts state transitions based on the input while maintaining linear-time complexity.

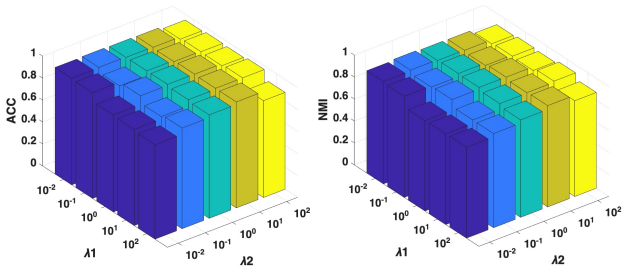

Figure 3: Influence of parameters λ_1 and λ_2 on clustering performance on HW.

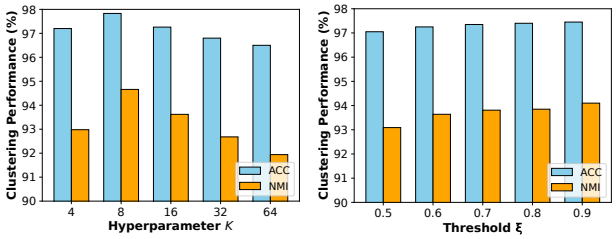

Figure 4: Influence of hyperparameter K and threshold ξ on clustering performance on HW.

Methods	Backbone	ACC (%)	Runtime (h)
MFLVC	AE	27.70	3.59
GCFagg	Transformer	<u>32.64</u>	7.53
DealMVC	AE	23.76	2.32
SEM	AE	30.39	2.93
TMCN	Mamba	31.30	<u>2.15</u>
GHICMC	GNN	32.43	4.26
MGLC	Mamba	**34.52**	**1.28**

Table 5: Comparison of ACC and runtime for different methods on YouTubeFace. Runtime is the total training time.

Conclusion

In this paper, we propose a novel end-to-end MVC framework, MGLC, which explicitly models cross-view dependencies and global-local structural relationships among samples to effectively enhance inter-cluster separability and intra-cluster compactness. By flexibly constructing multi-view sequences, MGLC fully leverages Mamba's efficient sequence modeling capabilities to jointly model cross-view dependencies and structural priors. Furthermore, we design a Cross-Mamba Fusion module to dynamically integrate cross-view and global-local structural representations. Additionally, we develop a Dual Calibration Contrastive Learning module that adaptively refines both feature and semantic representations, while mitigating false negatives among semantically similar samples. Extensive experiments on eight benchmark datasets demonstrate that MGLC outperforms state-of-the-art methods in both clustering performance and computational efficiency.

Acknowledgments

This work was supported by the Big Data Computing Center of Southeast University and by the Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China.

References

Asuncion, A.; and Newman, D. 2007. UCI machine learning repository.

Bauckhage, C. 2015. K-means clustering is matrix factorization. *arXiv preprint arXiv:1512.07548*.

Cai, X.; Wang, H.; Huang, H.; and Ding, C. 2012. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12): i16–i24.

Chao, G.; Xu, K.; Xie, X.; and Chen, Y. 2025. Global Graph Propagation with Hierarchical Information Transfer for Incomplete Contrastive Multi-view Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 15713–15721.

Geusebroek, J.-M.; Burghouts, G. J.; and Smeulders, A. W. 2005. The Amsterdam library of object images. *International Journal of Computer Vision*, 61: 103–112.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Gu, A.; Goel, K.; and Ré, C. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.

Guan, R.; Liu, T.; Tu, W.; Tang, C.; Luo, W.; and Liu, X. 2025a. Sampling Enhanced Contrastive Multi-View Remote Sensing Data Clustering with Long-Short Range Information Mining. *IEEE Transactions on Knowledge and Data Engineering*.

Guan, R.; Tu, W.; Wang, S.; Liu, J.; Hu, D.; Tang, C.; Feng, Y.; Li, J.; Xiao, B.; and Liu, X. 2025b. Structure-adaptive multi-view graph clustering for remote sensing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16933–16941.

Jiang, X.; He, B.; Zhou, P. Y.; Chen, X.; Guo, J.; Xu, J.; and Liao, Y. 2025. A Unified Framework to BRIDGE Complete and Incomplete Deep Multi-View Clustering under Non-IID Missing Patterns. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 594–603.

Kim, J.; and Park, H. 2024. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter conference on applications of computer Vision*, 7604–7613.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, H.; Li, Y.; Yang, M.; Hu, P.; Peng, D.; and Peng, X. 2023. Incomplete multi-view clustering via prototype-based imputation. *arXiv preprint arXiv:2301.11045*.

Li, Y.; Li, H.; Lin, Y.; Zhang, D.; Peng, D.; Liu, X.; Xie, J.; Hu, P.; Chen, L.; Luo, H.; et al. 2025. MetaQ: fast, scalable and accurate metacell inference via single-cell quantization. *Nature Communications*, 16(1): 1205.

Lin, Y.; Gou, Y.; Liu, X.; Bai, J.; Lv, J.; and Peng, X. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4447–4461.

Lin, Y.; Gou, Y.; Liu, Z.; Li, B.; Lv, J.; and Peng, X. 2021. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11174–11183.

Lin, Y.; Yang, M.; Yu, J.; Hu, P.; Zhang, C.; and Peng, X. 2023. Graph matching with bi-level noisy correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, 23362–23371.

Lin, Y.; Zhang, J.; Huang, Z.; Liu, J.; Peng, X.; et al. 2024. Multi-granularity Correspondence Learning from Long-term Noisy Videos. In *International Conference on Learning Representations*.

Lu, Y.; Li, H.; Li, Y.; Lin, Y.; and Peng, X. 2024. A survey on deep clustering: from the prior perspective. *Vicinagearth*, 1(1): 4.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pechlivanidou, G.; and Karampetakis, N. 2022. Zero-order hold discretization of general state space systems with input delay. *IMA Journal of Mathematical Control and Information*, 39(2): 708–730.

Ren, Y.; Pu, J.; Cui, C.; Zheng, Y.; Chen, X.; Pu, X.; and He, L. 2024. Dynamic weighted graph fusion for deep multi-view clustering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 4842–4850.

Sun, T.-K.; Chen, S.-C.; Jin, Z.; and Yang, J.-Y. 2007. Kernelized discriminative canonical correlation analysis. In *2007 International Conference on Wavelet Analysis and Pattern Recognition*, volume 3, 1283–1287. IEEE.

Sun, Y.; Li, Y.; Ren, Z.; Duan, G.; Peng, D.; and Hu, P. 2025. ROLL: Robust Noisy Pseudo-label Learning for Multi-View Clustering with Noisy Correspondence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 30732–30741.

Sun, Y.; Qin, Y.; Li, Y.; Peng, D.; Peng, X.; and Hu, P. 2024. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*.

Trosten, D. J.; Lokse, S.; Jenssen, R.; and Kampffmeyer, M. 2021. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1255–1265.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wan, X.; Xiao, B.; Liu, X.; Liu, J.; Liang, W.; and Zhu, E. 2024. Fast continual multi-view clustering with incomplete views. *IEEE Transactions on Image Processing*, 33: 2995–3008.

Wang, Q.; Zhang, Z.; Feng, W.; Tao, Z.; and Gao, Q. 2025. Contrastive Multi-view Subspace Clustering via Tensor Transformers Autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21207–21215.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *International conference on machine learning*, 1083–1092. PMLR.

Xing, S.; Qian, C.; Wang, Y.; Hua, H.; Tian, K.; Zhou, Y.; and Tu, Z. 2025. Openemma: Open-source multimodal model for end-to-end autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision*, 1001–1009.

Xu, J.; Chen, S.; Ren, Y.; Shi, X.; Shen, H.; Niu, G.; and Zhu, X. 2024. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. *Advances in Neural Information Processing Systems*, 36.

Xu, J.; Ren, Y.; Li, G.; Pan, L.; Zhu, C.; and Xu, Z. 2021. Deep embedded multi-view clustering with collaborative training. *Information Sciences*, 573: 279–290.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022. Multi-Level Feature Learning for Contrastive Multi-View Clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16051–16060.

Xu, J.; Zhao, N.; Niu, G.; Sugiyama, M.; and Zhu, X. 2025. Robust Multi-View Learning via Representation Fusion of Sample-Level Attention and Alignment of Simulated Perturbation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4232–4241.

Yan, W.; Zhang, Y.; Lv, C.; Tang, C.; Yue, G.; Liao, L.; and Lin, W. 2023. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19863–19872.

Yan, W.; Zhang, Y.; Tang, C.; Zhou, W.; and Lin, W. 2024. Anchor-Sharing and Clusterwise Contrastive Network for Multiview Representation Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Yang, M.; Li, Y.; Hu, P.; Bai, J.; Lv, J.; and Peng, X. 2022. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 1055–1069.

Yang, M.; Li, Y.; Huang, Z.; Liu, Z.; Hu, P.; and Peng, X. 2021. Partially view-aligned representation learning with noise-robust contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1134–1143.

Yang, X.; Jiaqi, J.; Wang, S.; Liang, K.; Liu, Y.; Wen, Y.; Liu, S.; Zhou, S.; Liu, X.; and Zhu, E. 2023. Dealmvc: Dual contrastive calibration for multi-view clustering. In *Proceedings of the 31st ACM International Conference on Multimedia*, 337–346.

Zhang, C.; Wang, Z.; Jia, X.; Li, Z.; Chen, C.; and Li, H. 2025a. Multi-view Clustering with Incremental Instances and Views. *IEEE Transactions on Image Processing*.

Zhang, Y.; Lin, Y.; Yan, W.; Yao, L.; Wan, X.; Li, G.; Zhang, C.; Ke, G.; and Xu, J. 2025b. Incomplete Multi-view Clustering via Diffusion Contrastive Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22650–22658.

Zhang, Y.; Yan, W.; Tang, C.; Zhou, W.; and Jin, J. 2025c. Multi-branch Space Sharing Feature Aggregation for contrastive multi-view clustering. *Pattern Recognition*, 111704.

Zhu, J.; Zou, X.; Liu, L.; Huang, Z.; Zhang, Y.; Tang, C.; and Dai, L.-R. 2025a. Trusted Mamba Contrastive Network for Multi-View Clustering. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*.

Zhu, Y.; Zheng, X.; He, X.; Zou, X.; Wang, P.; Tang, C.; Liu, X.; and He, K. 2025b. BMCST: Balanced multi-view clustering for spatially resolved transcriptomics with mamba-driven dynamic feature refinement. *Information Fusion*, 103425.