SCD-MVC: Stable Conditional Diffusion for Multi-view Clustering

Jinli Ma[1,*], **Chenkai Guo**[2,5,*], **Renda Han**[1], **Renxiang Guan**[3], **Siwei Wang**[4], **Ke Liang**[3], **Xiaoyu Cui**[1] and **Dayu Hu**[1,†]
[1]College of Medicine and Biological Information Engineering, Northeastern University
[2]Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences
[3]College of Computer Science and Technology, National University of Defense Technology
[4]Intelligent Game and Decision Lab, Academy of Military Sciences
[5]University of Chinese Academy of Sciences
jma54896@gmail.com, shaunak.guo@gmail.com, hudy@bmie.neu.edu.cn

Abstract

Multi-view clustering has garnered significant attention for its ability to integrate heterogeneous data and uncover underlying categorical structures. However, prevailing autoencoder-based methods often yield indiscriminative embeddings, rendering them prone to trivial solutions. While diffusion models have shown promise in modeling complex data distributions, their generative processes remain inherently unstable. More critically, the resulting latent representations often lack sufficient intra-class compactness and inter-class separability, leading to suboptimal clustering performance. In this paper, we propose SCD-MVC, a **S**table **C**onditional **D**iffusion-driven framework for **M**ulti-**V**iew **C**lustering. Specifically, it integrates conditional diffusion generation into the latent encoder to generate a target view conditioned on the remaining views, thereby facilitating robust joint distribution modeling. To align generation with clustering objectives, we introduce a semantic alignment module that steers the diffusion path toward view-consistent and discriminative representations. Furthermore, we introduce a consistency-aware regularization that enforces alignment between representations of adjacent diffusion steps. This stabilizes training and guides the optimization toward clustering-friendly representations. Extensive experiments on nine benchmark datasets demonstrate that SCD-MVC consistently outperforms eight state-of-the-art methods, achieving performance gains ranging from 6.68% to 13.09%. The source code is available at https://github.com/Knighttt0011/SCD-MVC.

1 Introduction

The rapid proliferation of diverse sensors and data acquisition modalities has produced vast amounts of multi-view data, which provide complementary representations of underlying entities[Li *et al.*, 2018b; 2018a; Yan *et al.*, 2021]. Consequently, multi-view clustering (MVC) has emerged as a pivotal unsupervised learning paradigm that partitions heterogeneous data by integrating consistent patterns across multiple views [Fang *et al.*, 2025]. By effectively leveraging cross-view synergies, MVC has demonstrated strong applicability across a wide range of domains, including computer vision, social network analysis, and bioinformatics[Lu *et al.*, 2024; Li *et al.*, 2023; Sun *et al.*, 2024].

Traditional MVC methodologies, such as multi-view spectral clustering and canonical correlation analysis, primarily focus on modeling shallow linear dependencies and simple geometric structures [Fu *et al.*, 2020; Yang *et al.*, 2025]. Driven by advances in deep learning, deep MVC has emerged as a dominant paradigm, leveraging autoencoders or contrastive frameworks to extract nonlinear, high-level semantic features [Tang *et al.*, 2022; Xu *et al.*, 2022; Fei *et al.*, 2025; Ding *et al.*, 2025]. While these discriminative methods have achieved remarkable success, they typically rely on deterministic mappings optimized through reconstruction or contrastive losses. Crucially, these methods often overlook the underlying generative data distribution. This limitation frequently leads to suboptimal representation expressiveness and increased susceptibility to trivial solutions (e.g., mode collapse), particularly in the presence of complex noise or substantial inter-view discrepancies[Xi *et al.*, 2025].

Diffusion Probabilistic Models (DPMs) have recently revolutionized generative learning by achieving unprecedented sample fidelity through iterative denoising processes [Ho *et al.*, 2020]. Despite their exceptional capacity for learning expressive representations, adapting DPMs to multi-view clustering poses two non-trivial challenges: (1) **Diffusion Instability.** Standard diffusion models are inherently stochastic in nature[Kang *et al.*, 2021]. The randomness embedded in the diffusion trajectory often induces significant fluctuations in latent embeddings across diffusion timesteps. Such temporal inconsistency hinders the formation of compact and discriminative clusters [Biroli *et al.*, 2024; Smith and Dalchau, 2018]; (2) **Gap between Clustering and Generation.** There exists a fundamental mismatch between generative capability and clustering requirements[Gunawardena *et al.*, 2024]. While DPMs can effectively generate high-quality data representations, their generative objectives do not explicitly enforce discriminative separability in the learned latent space, which is

essential for clustering and, if unaddressed, may lead to overlapping cluster boundaries [Li *et al.*, 2025].

To address these challenges, we propose **SCD-MVC**, a novel **S**table **C**onditional **D**iffusion framework for **M**ulti-**V**iew **C**lustering. Instead of treating view generation in isolation, we reformulate the problem as a conditional generation task, in which each target view is synthesized conditioned on the context provided by all remaining views. This formulation explicitly encourages the latent space to aggregate complementary cross-view information. To mitigate the inherent instability of the diffusion process, we introduce a consistency-aware regularization scheme that enforces smoothness across representation trajectories between adjacent diffusion steps. Furthermore, to bridge the misalignment between generative objectives and clustering targets, we design a cluster-guided semantic fusion module. By leveraging a Transformer architecture to integrate cross-view features and refining the latent space through iterative pseudo-labeling, this module ensures the learned representations are both geometrically compact and highly discriminative. The main contributions of this paper are summarized as follows:

- *Framework:* We propose SCD-MVC, a stable generative multi-view clustering framework that synergizes conditional diffusion with clustering assignments. To the best of our knowledge, this represents one of the pioneering works to address diffusion stability within the context of clustering-guided representation generation.
- *Algorithm:* We introduce a consistency-aware regularization that constrains the denoising network to yield consistent predictions across adjacent time steps t and $t-1$, thereby imposing smoothness on the induced generative manifold. Furthermore, we develop a cluster-guided semantic fusion mechanism to align the generative process with the discriminative clustering objective.
- *Evaluation:* Extensive experiments on nine benchmark datasets demonstrate that SCD-MVC significantly outperforms state-of-the-art methods, achieving an average improvement of up to 13.09% in clustering accuracy.

2 Related Work

2.1 Multi-view Clustering

MVC aims to partition data into distinct groups by integrating heterogeneous information from diverse views [Fang *et al.*, 2023]. Early research primarily relied on matrix factorization and spectral analysis to uncover shared latent structures while addressing noise and scalability [Fu *et al.*, 2020]. Notably, RMKMC proposes a robust framework designed for large-scale data, using structured sparsity to integrate heterogeneous features while effectively mitigating the impact of outliers[Cai *et al.*, 2013]. Similarly, to capture the topological structure of data, MSGL introduces a multi-view spectral graph learning strategy that optimizes the graph Laplacian matrix to preserve consistency across data from different views [Kang *et al.*, 2021]. With the rapid development of deep learning, methods have evolved to learn more expressive non-linear representations [Chen *et al.*, 2022]. DCCNMF combined deep learning with matrix factorization, enforcing consensus across views through a deep canonical architecture [Gunawardena *et al.*, 2024]. Focusing on efficiency, FastMICE significantly reduces computational overhead by employing a random-forest-based ensemble strategy to synthesize consensus cross-view representations [Huang *et al.*, 2023]. In the field of graph learning, GCFAggMVC introduces a graph collaborative filtering aggregation mechanism, which effectively captures high-order structural information and non-linear interactions across views [Yan *et al.*, 2023].

Most recently, contrastive and regression-based paradigms have achieved state-of-the-art performance by aligning view-specific distributions. CCMVC leverages context-aware contrastive learning to align samples at both the instance and cluster levels [Shi *et al.*, 2025]. Addressing the challenge of incomplete data, DealMVC tackles view misalignment through a dual-contrastive prediction mechanism that recovers missing semantics [Yang *et al.*, 2023]. Furthermore, CRMVC presents a continuum regression framework, flexibly bridging canonical correlation analysis and partial least squares to extract robust common representations [Kumar *et al.*, 2011]. Notwithstanding these advancements, the majority of existing deep MVC methods are still predominantly built upon autoencoder-based backbones.[Zhou *et al.*, 2024] However, such approaches often produce embeddings with limited discriminative power and fail to explicitly model the underlying generative distribution of complex multi-view data.[Zhang *et al.*, 2024]

2.2 Generative Diffusion Model

Diffusion Probabilistic Models, encompassing Denoising Diffusion Probabilistic Models (DDPMs) and Score-based Generative Models (SGMs), have emerged as a powerful class of generative frameworks [Ho *et al.*, 2020; Song *et al.*, 2020; Croitoru *et al.*, 2023]. Unlike generative adversarial networks or VAEs, diffusion models generate high-fidelity samples by reversing a gradual noise-adding process via a learnable Markov chain[Graham *et al.*, 2022]. The core objective involves optimizing a variational lower bound or matching the score function of the data distribution.

Recent advancements, exemplified by Latent Diffusion Models and ControlNet, have extended diffusion models to conditional generation and representation learning. Conditional Diffusion Models incorporate auxiliary information (e.g., class labels or text prompts) into the denoising process to guide the synthesis towards desired attributes [Rombach *et al.*, 2022; Zhang *et al.*, 2023; Croitoru *et al.*, 2023]. In the context of unsupervised learning, diffusion models are increasingly explored for their ability to learn separate latent structures [Croitoru *et al.*, 2023]. However, applying conditional diffusion mechanisms to the MVC domain remains largely under-explored. Existing generative clustering methods often fail to explicitly model the conditional dependencies between views, which induces significant fluctuations in latent embeddings across diffusion timesteps and thus hinders the formation of compact and discriminative clusters.

3 Methods

3.1 Problem Definition

Let $\mathcal{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{V}$ denote a multi-view dataset with V views, where $\mathbf{X}^{(v)} = [\mathbf{x}_1^{(v)}, \ldots, \mathbf{x}_N^{(v)}]^\top \in \mathbb{R}^{N \times d_v}$ represents the data matrix for the v-th view with N samples and d_v dimensions. The goal of MVC is to partition the N samples into K disjoint clusters $\mathcal{C} = \{C_1, \ldots, C_K\}$. Formally, MVC learns a mapping function $f : \mathcal{X} \rightarrow \mathbf{y}$ that assigns a cluster label $y_i \in \{1, \ldots, K\}$ to the i-th sample. Unlike traditional methods that cluster raw features directly, we reformulate the task as learning a generative latent embedding using conditional diffusion. The proposed approach jointly optimizes the generative likelihood of view-specific data $p(\mathbf{X}^{(v)} \mid \mathbf{Z})$ and the discriminative power of the latent representation $\mathbf{Z}$. As a result, samples with the same semantics are consistently aggregated across heterogeneous views.

3.2 View-specific Latent Embedding

The high dimensionality and heterogeneity of multi-view data make direct generative modeling challenging. Following common practice, we project features from different views into a compact latent space using view-specific autoencoders. Specifically, for the v-th view, an encoder $f_{\theta^{(v)}}^{(v)}(\cdot)$ maps an input sample $\mathbf{x}_i^{(v)}$ to a latent embedding $\mathbf{z}_i^{(v)}$, while the corresponding decoder $g_{\eta^{(v)}}^{(v)}(\cdot)$ reconstructs the input to preserve semantic information. Formally, the encoding and decoding processes are defined as follows:

$$\hat{\boldsymbol{x}}_i^{(v)} = g_{\eta^{(v)}}^{(v)}(\boldsymbol{z}_i^{(v)}) = g_{\eta^{(v)}}^{(v)}\left(f_{\theta^{(v)}}^{(v)}(\boldsymbol{x}_i^{(v)})\right). \tag{1}$$

To ensure that $\mathbf{z}_i^{(v)}$ preserves the essential semantic information of the original data, we optimize the representations by minimizing the mean squared error (MSE) between the original inputs and the reconstructed outputs:

$$\mathcal{L}_{\text{res}} = \sum_{v=1}^{V} \sum_{i=1}^{N} ||\boldsymbol{x}_i^{(v)} - g_{\eta^{(v)}}^{(v)}\left(f_{\theta^{(v)}}^{(v)}(\boldsymbol{x}_i^{(v)})\right)||_2^2, \tag{2}$$

this process aligns diverse feature spaces into a unified dimension d_h, providing a robust and compact representation for the subsequent conditional diffusion generation.

3.3 Stable Conditional Diffusion Generation

Diffusion probabilistic models have recently demonstrated strong capabilities in generative representation learning. To explicitly model intricate correlations across heterogeneous views and promote a consensus data representation, we construct a conditional diffusion probabilistic model in the learned latent space. Unlike standard unconditional diffusion paradigms, the proposed module is specifically designed for the MVC setting, in which it generates the latent representation of a target view by conditioning on semantic context extracted from the remaining complementary views.

Cross-view Conditional Diffusion. Formally, the forward diffusion process is modeled as a fixed Markov chain that progressively corrupts the clean latent representation $\boldsymbol{z}_{i,0}^{(v)}$ with Gaussian noise. This procedure generates a sequence of noisy latents $\{\boldsymbol{z}_{i,t}^{(v)}\}_{t=1}^{T}$, regulated by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$. The transition kernel at timestep t is formulated as:

$$q(\boldsymbol{z}_{i,t}^{(v)} | \boldsymbol{z}_{i,t-1}^{(v)}) = \mathcal{N}(\boldsymbol{z}_{i,t}^{(v)}; \sqrt{1-\beta_t}\boldsymbol{z}_{i,t-1}^{(v)}, \beta_t \mathbf{I}), \tag{3}$$

where $\mathbf{I}$ denotes the identity matrix and $\mathcal{N}$ represents the normal distribution. To enhance the signal-to-noise ratio during the diffusion process, we employ a cosine schedule for β_t, which effectively mitigates the abrupt information collapse often observed in linear schedules. Leveraging the closed-form property of the Gaussian distribution, we can sample $\boldsymbol{z}_{i,t}^{(v)}$ at an arbitrary timestep t directly from $\boldsymbol{z}_{i,0}^{(v)}$:

$$\boldsymbol{z}_{i,t}^{(v)} = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_{i,0}^{(v)} + \sqrt{1-\bar{\alpha}_t}\epsilon, \;\; \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s), \tag{4}$$

here, ϵ denotes the noise term and is assumed to follow a standard Gaussian distribution, i.e., $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In addition, we adopt an offset noise strategy as a lightweight stabilization heuristic, following common practice in diffusion-based models. Specifically, the Gaussian noise is perturbed by a learnable channel-wise offset:

$$\xi = \epsilon + \lambda_{\text{off}} \cdot \delta, \;\; \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

where λ_{off} controls the magnitude of the offset. The reverse diffusion process aims to recover the latent variable $\boldsymbol{z}_{i,0}^{(v)}$ by progressively removing the injected noise. We formulate this reverse process as a conditional denoising task, in which the generation of the v-th view is guided by the semantic context of all complementary views $\{1, 2, \ldots, v-1, v+1, \ldots, V\}$. We construct a conditioning vector $\boldsymbol{c}_i^{(v)}$ by concatenating the latent representations of the remaining views as follows:

$$\boldsymbol{c}_i^{(v)} = \bigoplus_{u \neq v} \boldsymbol{z}_{i,0}^{(u)}, \tag{6}$$

where the $\bigoplus$ denotes the concatenation operation. This conditioning approach is designed for standard low-view MVC benchmarks, where the number of views is small and the resulting dimensionality remains tractable. The denoising diffusion network $\epsilon_\phi^{(v)}$ takes the noisy latent $\boldsymbol{z}_{i,t}^{(v)}$, the time embedding $\tau(t)$, and the conditioning vector $\boldsymbol{c}_i^{(v)}$ as inputs to predict the noise component. The sinusoidal time embedding $\tau(t)$ injects temporal position information and is defined as $\tau(t) = \left[\ldots, \sin(t/10000^{2k/d}), \cos(t/10000^{2k/d}), \ldots\right]$.

We train the diffusion model using a hybrid objective that balances reconstruction fidelity and temporal coherence. This objective consists of a denoising term and a consistency-aware regularizer. Specifically, the denoising loss $\mathcal{L}_{\text{denoise}}$ minimizes the discrepancy between the target noise ξ and the predicted noise component:

$$\mathcal{L}_{\text{denoise}} = \sum_{v=1}^{V} \mathbb{E}_{t, \boldsymbol{z}_{i,0}^{(v)}, \xi}\left[||\xi - \epsilon_\phi^{(v)}(\boldsymbol{z}_{i,t}^{(v)}, \tau(t), \mathbf{c}_i^{(v)})||_2^2\right], \tag{7}$$

this term guides the model to learn the reverse transition kernel required to reconstruct the original data distribution.

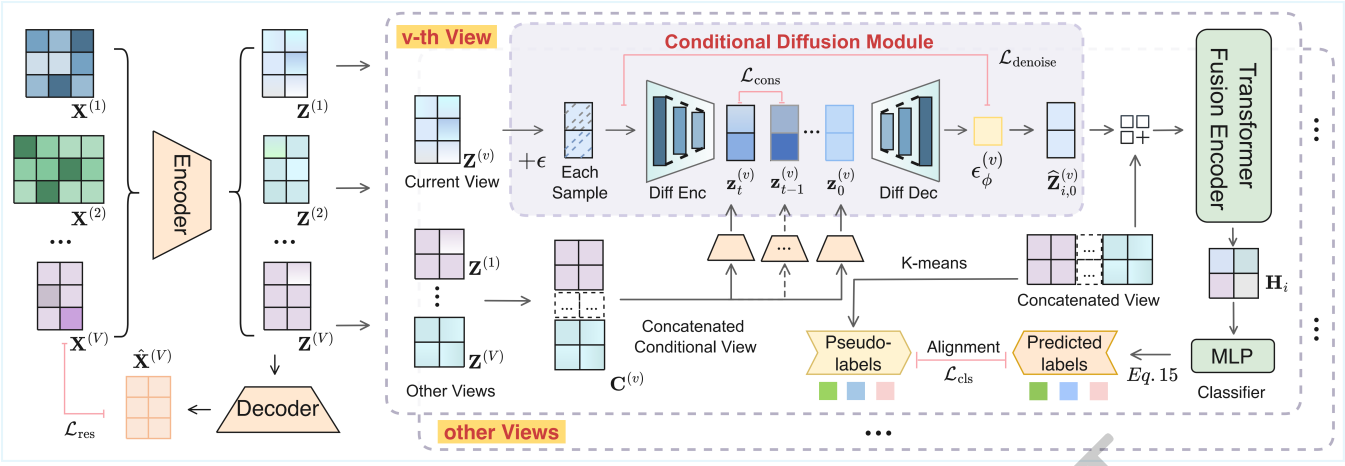


Figure 1: Overview of the SCD-MVC framework. It integrates three main modules: (1) view-specific encoder, which projects raw features into compact latent embeddings via autoencoders; (2) stable cross-view conditional diffusion, which learns consensus representations by generating target-view latents conditioned on complementary view contexts and stabilized through consistency optimization; and (3) clustering-guided semantic alignment, which fuses denoised representations using a Transformer and aligns them with global centroids via iterative pseudo-labeling.

Consistency-aware Optimization. However, the inherent stochasticity of the diffusion process often leads to significant fluctuations during sampling. To impose smoothness on the induced generative manifold, we introduce a consistency-aware regularization strategy. Specifically, we enforce temporal coherence by constraining the denoising network to yield consistent predictions for the same sample across adjacent time steps t and $t-1$. Let $\boldsymbol{z}_{i,t-1}^{(v)}$ represent the noisy latent at step $t-1$, derived from the same clean latent $\boldsymbol{z}_{i,0}^{(v)}$ and noise ξ. The consistency loss is formulated as:

$$\mathcal{L}_{\text{cons}} = \sum_{v=1}^{V} \mathbb{E}_{t>1} \left[\left\| \epsilon_\phi^{(v)}(\boldsymbol{z}_{i,t}^{(v)}, t, \mathbf{c}_i^{(v)}) - \epsilon_\phi^{(v)}(\boldsymbol{z}_{i,t-1}^{(v)}, t-1, \mathbf{c}_i^{(v)}) \right\|_2^2 \right], \tag{8}$$

this regularization effectively constrains the gradient of the learned vector field, ensuring that the semantic trajectory remains stable throughout the reverse diffusion process. Finally, the total objective combines the standard denoising loss with the consistency term, balanced by a hyperparameter α:

$$\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{denoise}} + \alpha \mathcal{L}_{\text{cons}}. \tag{9}$$

Empirically, we set $\alpha = 50$ to strike a balance between training stability and denoising precision. This configuration fosters the learning of robust cross-view correlations without compromising generation quality. Through the conditional diffusion process, the objective enforces the latent representation of a target view to be both predictive of and consistent with the remaining views.

3.4 Clustering-guided Semantic Alignment

While the conditional diffusion module effectively generates high-quality embeddings, it does not explicitly guarantee that the learned latent space exhibits the discriminative separability required for clustering. To bridge the gap between the generative capability of diffusion models and the discrete nature of clustering tasks, we introduce a novel cluster-guided semantic fusion mechanism driven by iterative pseudo-labeling. The core motivation is to leverage the intrinsic geometric structure of the fused latent representations as semantic anchors, thereby guiding the view-specific encoders and the fusion transformer to pull intra-class samples closer while pushing inter-class samples further apart. This module extracts a denoised feature representation and optimizes it with respect to global cluster centroids.

Semantic Fusion via Transformer. To initiate the process, we recover the clean latent representation $\hat{\boldsymbol{z}}_{i,0}^{(v)}$ from the current noisy state $\boldsymbol{z}_{i,t}^{(v)}$ by leveraging the predicted noise:

$$\hat{\boldsymbol{z}}_{i,0}^{(v)} = \frac{1}{\sqrt{\bar{\alpha}t}} \left(\boldsymbol{z}_{i,t}^{(v)} - \sqrt{1-\bar{\alpha}t}\, \epsilon_\phi^{(v)}(\boldsymbol{z}_{i,t}^{(v)}, t, \mathbf{c}_i^{(v)}) \right), \tag{10}$$

$\hat{\boldsymbol{z}}_{i,0}^{(v)}$ acts as a proxy for the view-specific latent representation. The asymmetric token construction reflects the unequal reliability between the recovered target representation and clean complementary-view representations. The latter provides reliable semantic context to refine the uncertain target token, following the reconstructive learning principle that clean context can guide target recovery [He *et al.*, 2022]. To synthesize cross-view information for holistic clustering, we formulate a sequence S_i that incorporates both the recovered target view and the corresponding clean complementary views:

$$S_i = \left[\hat{\boldsymbol{z}}_{i,0}^{(v)}, \boldsymbol{z}_{i,0}^{(1)}, \ldots, \boldsymbol{z}_{i,0}^{(v-1)}, \boldsymbol{z}_{i,0}^{(v+1)}, \ldots, \boldsymbol{z}_{i,0}^{(V)} \right], \tag{11}$$

then, the sequence is processed by a Transformer encoder, which leverages self-attention mechanisms to capture high-order cross-view dependencies. The output embedding corresponding to the target view token constitutes the fused semantic representation $\mathbf{h}_i^{(v)}$, formulated as:

$$\mathbf{H}_i = \text{TransformerEncoder}(S_i); \quad \mathbf{h}_i^{(v)} = \mathbf{H}_i[0], \tag{12}$$

$\mathbf{H}_i$ encapsulates the output embeddings of the sequence. We specifically designate the first token $\mathbf{H}_i[0]$ as the fused representation $\mathbf{h}_i^{(v)}$, as it collects and synthesizes information from all views through the attention mechanism.

Iterative Pseudo-Labeling. To obtain the final cluster assignments, we utilize a classifier $f_{\text{cls}}(\cdot)$ to map $\mathbf{h}_i^{(v)}$ to the target category distribution. The predictive probability of sample i being assigned to cluster k is derived as follows:

$$p_{ik}^{(v)} = \frac{\exp\left(\left(f_{cls}(\mathbf{h}_i^{(v)})\right)_k\right)}{\sum_{j=1}^{K}\exp\left(\left(f_{cls}(\mathbf{h}_i^{(v)})\right)_j\right)}. \tag{13}$$

In parallel, we maintain a set of global cluster centroids $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$ within the joint latent space. These centroids are updated iteratively to align with the distribution of the fused representations. We then derive pseudo-labels $\hat{y}_i$ for each sample by assigning it to the nearest centroid based on feature proximity:

$$\hat{y}_i = \arg\min_{k\in 1,\ldots,K} \left\| \bigoplus_{v=1}^{V} \boldsymbol{z}_{i,0}^{(v)} - \mu_k \right\|_2^2. \tag{14}$$

To enforce clustering consistency, we minimize the Cross-Entropy loss between the predicted soft assignments and the target pseudo-labels. This objective effectively promotes intra-cluster compactness by guiding samples towards their assigned centroids:

$$\mathcal{L}_{\text{cls}} = \sum_{v=1}^{V}\sum_{k=1}^{K} \mathbb{I}(k = \hat{y}_i)\log(p_{ik}^{(v)}), \tag{15}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. This aligning loss effectively encourages the latent embeddings to cluster tightly around their corresponding centroids across all V views, thereby producing clustering-friendly embeddings.

3.5 Joint Optimization

To enable mutually beneficial learning across different tasks, we optimize our SCD-MVC in an end-to-end manner by minimizing a unified objective function. This objective integrates the reconstruction, diffusion, and clustering losses as follows:

$$\mathcal{L} = \mathcal{L}_{\text{res}} + \lambda_1\mathcal{L}_{\text{diff}} + \lambda_2\mathcal{L}_{\text{cls}}, \tag{16}$$

where λ_1 and λ_2 are trade-off hyperparameters that balance the contributions of reconstruction fidelity, generative capability, and clustering discriminability, respectively.

4 Experiments

4.1 Experimental Settings

Benchmark Datasets. To evaluate the effectiveness of the proposed SCD-MVC method, we conduct extensive experiments on nine diverse and widely used multi-view benchmark datasets, including YaleA, Hdigit, Dermatology, Prokaryotic, Washington, ORL-3Views, ForestTypes, WebKB, and BBC-Sport. Following standard experimental protocols, all feature vectors are standardized to have zero mean and unit variance.

Dataset	n	v	k	d
YaleA	165	3	15	512/50/9
Washington	230	4	5	230/1703/230/230
Dermatology	358	2	6	12/22
ORL 3Views	400	3	40	3304/6750/4096
ForestTypes	523	3	4	9/9/9
BBCSport	544	2	5	3203/3183
Prokaryotic	551	2	4	393/438/3
WebKB	1051	2	2	2949/334
Hdigit	10000	2	10	784/256

Table 1: Nine multi-view benchmark datasets.

No additional preprocessing steps are applied to ensure a fair comparison among different methods.

Evaluation Metrics. To rigorously evaluate clustering performance, we adopt three standard evaluation metrics: Accuracy (ACC), Normalized Mutual Information (NMI), and Purity (PUR). Following established evaluation protocols, the ACC score is computed by identifying the optimal permutation that aligns the predicted cluster assignments with the ground-truth labels using the Kuhn–Munkres algorithm.

Baseline Methods. Our comparative study includes eight widely cited and recent baseline methods: CRMVC [Kumar *et al.*, 2011], RMKMC [Cai *et al.*, 2013], MSGL [Kang *et al.*, 2021], DealMVC [Yang *et al.*, 2023], GCFAggMVC [Yan *et al.*, 2023], FastMICE [Huang *et al.*, 2023], DCCNMF [Shi *et al.*, 2025], and CCMVC [Shi *et al.*, 2025].

Implementation Details. Our method is implemented in PyTorch and trained on NVIDIA RTX 3090 GPUs. The model is trained for up to 200 epochs with an early stopping strategy to prevent overfitting. To ensure a fair comparison, we report the best performance achieved by all baseline methods. For datasets not included in the original studies, results are reproduced using the authors' official code releases, with a grid search conducted within the recommended parameter ranges to obtain the best performance.

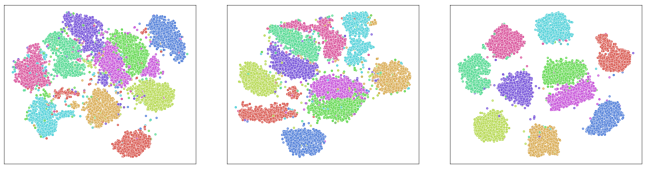
4.2 Clustering Performance

Table 2 presents the clustering results on nine benchmark datasets. The best results are highlighted in bold, while the second-best results are underlined. Based on these results, we draw the following conclusions:

(1) SCD-MVC consistently outperforms state-of-the-art baseline methods on the majority of datasets, achieving notable performance gains over the second-best competitors. For example, on the Dermatology dataset, SCD-MVC exceeds the second-best method by more than 13% in terms of ACC.

(2) SCD-MVC demonstrates a clear performance advantage on challenging datasets (e.g., Prokaryotic and ForestTypes) and exhibits strong scalability to large-scale scenarios (e.g., WebKB and Hdigit). These gains can be attributed to the proposed cross-view conditional diffusion framework, which captures fine-grained semantic correlations and facilitates robust joint representation learning across diverse and high-dimensional views.

(3) SCD-MVC achieves a superior balance between efficiency and accuracy. As shown in Table 3, while lightweight baselines trade performance for speed and



(a) w/o CV-Diff (b) w/o CG-Fuse (c) SCD-MVC

Figure 2: t-SNE visualization of latent representations on the Hdigit dataset. (a) w/o CV-Diff, (b) w/o CG-Fuse, (c) the proposed SCD-MVC. The complete model shows superior cluster separability.

complex models like CCMVC incur prohibitive overheads, SCD-MVC maintains a competitive runtime. Notably, our method is significantly faster than CCMVC, confirming that its substantial performance gains do not impose excessive computational burdens.

4.3 Ablation Study

Table 4 reports the ablation study results on the ForestTypes and Hdigit datasets, validating the effectiveness of key components. For brevity, we denote the Cross-View Conditional Diffusion and Clustering-Guided Semantic Fusion modules as CV-Diff and CG-Fuse, respectively.

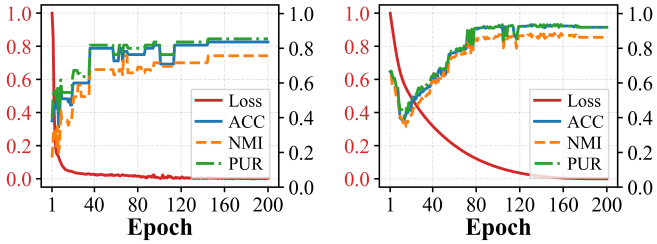
w/o CV-Diff: By removing the CV-Diff module, this variant eliminates the reverse reconstruction mechanism and loses the generative stable diffusion. Thus, performance degrades significantly on the spectrally overlapped *ForestTypes* dataset, highlighting the critical role of CV-Diff in maintaining cross-view robustness. In its absence, the latent spaces across different views remain misaligned, leading to increased semantic divergence and degraded embedding quality.

w/o CG-Fuse: By removing the CG-Fuse module, this variant becomes less discriminative, as the diffusion process alone cannot compress intra-class variance or enlarge inter-class margins. This leads to ambiguous K-means decision boundaries that incorrectly group heterogeneous samples. These results underscore that generative consistency alone, without clustering-driven discriminative guidance, fails to produce a compact and well-separated decision geometry.

To qualitatively analyze learned representations, we visualize feature distributions via t-SNE (Fig. 2). Both w/o CV-Diff and w/o CG-Fuse variants exhibit dispersed, overlapping clusters, indicating that reconstruction constraints or discriminative guidance alone are insufficient. Conversely, the full Model produces compact embeddings with clear margins, confirming that CV-Diff and CG-Fuse act synergistically to optimize clustering performance.

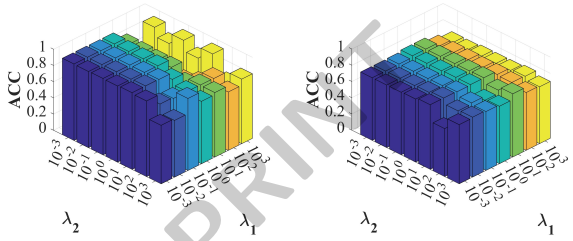
4.4 Convergence Analysis

To illustrate the optimization behavior of the proposed framework, we track the training dynamics of the normalized loss and key clustering metrics on the ForestTypes and Hdigit datasets. As illustrated in Fig. 3, the normalized loss decreases monotonically and converges near zero, while the clustering metrics show a steady upward trajectory, eventually saturating at optimal values without significant fluctuations. This consistent convergence behavior confirms that



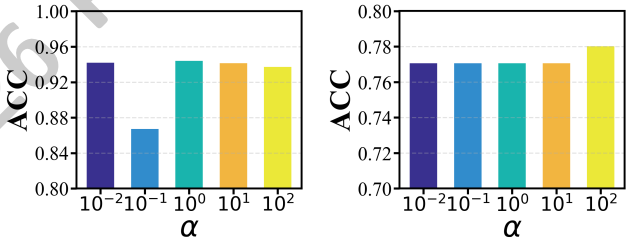
(a) ForestTypes (b) Hdigit

Figure 3: Convergence analysis on the ForestTypes and Hdigit datasets. The left Y-axis represents the Normalized Loss (red mark), while the right Y-axis denotes the clustering performance.



(a) ForestTypes (b) Hdigit

Figure 4: Sensitivity analysis of hyperparameters λ_1 and λ_2 on the ForestTypes and Hdigit datasets.



(a) ForestTypes (b) Hdigit

Figure 5: Sensitivity analysis of hyperparameters α on the ForestTypes and Hdigit datasets.

the synergy between CV-Diff and CG-Fuse effectively stabilizes the optimization process. The results demonstrate that our dual mechanism design, which balances generative constraints with discriminative regularization, ensures robust training stability across datasets of varying complexity. This also indicates that SCD-MVC achieves stable end-to-end joint optimization without relying on stage-wise pre-training or stop-gradient operations.

4.5 Parameter Analysis

To evaluate the robustness of the proposed framework, we conduct parameter sensitivity experiments, as illustrated in Fig. 4 and Fig. 5. We first analyze the sensitivity of the trade-off weights λ_1 and λ_2. The model maintains stable ACC across a wide range of $[10^{-3}, 10^{3}]$ on both the Hdigit and ForestTypes datasets. This observation indicates our method is highly tolerant to parameter perturbations. With optimal λ_1 and λ_2 values, we vary α within the range of $[10^{-2}, 10^{2}]$,

Dataset	CRMVC'11	RMKMC'13	MSGL'21	DealMVC'23	GCFAggMVC'23	FastMICE'23	DCCNMF'24	CCMVC'25	Ours
ACC									
YaleA	0.6148	0.6545	0.7505	0.3333	0.2970	0.6800	0.4078	0.5188	**0.7758**
Washington	0.4826	0.5217	0.5240	0.3174	0.3435	0.5621	0.3213	0.3802	**0.5957**
Dermatology	0.5123	0.7486	0.7084	0.7877	0.8017	0.7259	0.7374	0.6844	**0.9358**
ORL 3Views	0.6425	0.5275	0.6025	0.1025	0.2125	**0.7415**	0.3687	0.3438	0.6850
ForestTypes	0.7744	0.5143	0.7372	0.4302	0.3574	0.6772	0.5525	0.3574	**0.8203**
BBCSport	0.5000	0.5515	0.6526	0.6783	0.4412	0.3976	0.3654	0.3203	**0.7408**
Prokaryotic	0.5953	0.4356	0.6124	0.5336	0.4955	0.5537	0.4938	0.4336	**0.7650**
WebKB	0.5652	0.5186	0.6019	0.5119	0.5119	**0.7936**	0.7645	0.7090	0.7888
Hdigit	0.6853	0.6476	0.6868	0.9000	0.9816	0.9082	0.9446	**0.9872**	0.9467
Average	0.5969	0.5689	0.6529	0.5105	0.4936	0.6711	0.5507	0.5261	**0.7838**↑13.09%
NMI									
YaleA	0.6298	0.6880	0.6893	0.4448	0.4159	0.7235	0.4741	0.5790	**0.7838**
Washington	0.2085	0.1857	**0.3009**	0.0962	0.1167	0.2707	0.0463	0.1139	0.2023
Dermatology	0.5782	0.7110	0.7150	0.8094	0.7178	0.7853	0.7793	0.7334	**0.9040**
ORL 3Views	0.8444	0.7541	0.7999	0.3596	0.4698	**0.8915**	0.5902	0.6010	0.8039
ForestTypes	0.5290	0.3888	0.5232	0.1542	0.0956	0.5048	0.3695	0.0956	**0.5805**
BBCSport	0.3199	0.3846	0.4950	**0.5749**	0.2340	0.1288	0.1985	0.0546	0.4797
Prokaryotic	0.2703	0.2070	0.3959	0.2962	0.2746	0.2981	0.2438	0.2452	**0.4780**
WebKB	0.1173	0.0010	0.0103	0.1201	0.1121	0.0571	0.0052	**0.2426**	0.1513
Hdigit	0.6637	0.8024	0.7419	0.9375	0.9485	0.9208	0.9017	**0.9783**	0.8885
Average	0.4623	0.4581	0.5190	0.4214	0.3761	0.5090	0.4009	0.4048	**0.5858**↑6.68%
PUR									
YaleA	0.5758	0.6545	0.6236	0.3455	0.3333	0.6939	0.4327	0.5500	**0.7758**
Washington	0.6522	0.6261	0.6745	0.5304	0.5478	**0.6765**	0.5026	0.5885	0.6348
Dermatology	0.6508	0.7600	0.7808	0.7905	0.8017	0.8187	0.8296	0.8312	**0.9358**
ORL 3Views	**0.8509**	0.5925	0.6502	0.1025	0.2300	0.7767	0.3937	0.3724	0.6950
ForestTypes	0.7744	0.6386	0.7572	0.4688	0.4238	0.7246	0.6615	0.4238	**0.8203**
BBCSport	0.6250	0.6011	0.7067	**0.7868**	0.5037	0.4362	0.4693	0.3848	0.7408
Prokaryotic	0.6479	0.5717	0.7684	0.7169	0.6788	0.6562	0.3937	0.6016	**0.8330**
WebKB	0.7812	0.7812	0.7811	0.7812	0.7812	**0.7936**	0.7811	0.7812	0.7888
Hdigit	0.7189	0.8459	0.7795	0.9000	**0.9816**	0.9208	0.9446	0.9728	0.9567
Average	0.6975	0.6746	0.7247	0.6025	0.5869	0.7219	0.6010	0.6118	**0.7979**↑7.32%

Table 2: Clustering performance comparison on nine benchmark datasets. The best results are shown in **bold**, while the runners-up are underlined. ↑ denotes the improved performance compared with the second-best baseline.

Method	ForestTypes	Hdigit	100Leaves	prokaryotic	Washington	WebKB
CRMVC	13.14	871.38	919.04	15.53	6.75	97.02
MSGL	16.54	471.35	430.28	13.39	14.25	22.54
DealMVC	9.25	133.72	97.63	12.67	13.78	18.58
GCFAggT	22.45	348.42	63.58	13.55	11.66	54.35
FastMICE	18.15	222.66	123.54	28.96	12.51	204.95
DCCNMF	15.39	206.36	107.59	29.17	39.14	80.35
RMKM	12.66	139.42	100.12	36.44	26.56	97.80
CCMVC	247.67	4293.37	1512.44	1180.18	390.12	143.15
SCD-MVC	48.82	353.59	517.29	42.13	37.91	46.56

Table 3: Running time comparison of different baseline methods (in seconds).

Dataset	Modules	Metrics		
		ACC	NMI	PUR
ForestTypes	w/o CV-Diff	0.7706	0.5027	0.7706
	w/o CG-Fuse	0.6970	0.4884	0.7152
	SCD-MVC (Ours)	**0.8203**	**0.5805**	**0.8203**
Hdigit	w/o CV-Diff	0.9422	0.8793	0.9422
	w/o CG-Fuse	0.9421	0.8689	0.9421
	SCD-MVC (Ours)	**0.9467**	**0.8885**	**0.9567**

Table 4: Ablation study on the ForestTypes and Hdigit datasets. We compare the full model against two variants: (1) w/o CV-Diff and (2) w/o CG-Fuse, to validate the contribution of each component.

as shown in Fig. 5a and Fig. 5b. The model exhibits strong robustness. For example, ACC consistently exceeds 0.85 on Hdigit, while performance on ForestTypes follows a smooth curve without abrupt drops. These results further demonstrate the stability of the proposed design and the ease of hyperparameter tuning. Given this stability, we empirically set $\lambda_1 = 0.1$ and $\lambda_2 = 0.01$ for the main experiments.

5 Conclusions

In this paper, we propose SCD-MVC, a novel generative paradigm that reconsiders multi-view clustering through the lens of stable conditional diffusion. By explicitly modeling the joint distribution of heterogeneous views through cross-view conditional generation, our approach effectively captures complex latent semantics that are often overlooked by traditional discriminative methods. To mitigate the stochastic volatility inherent in diffusion probabilistic models, we introduce a consistency-aware regularization mechanism that ensures smooth and semantically consistent representation trajectories during the reverse diffusion process. Additionally, the proposed cluster-guided semantic fusion module bridges the objective gap between generative reconstruction and discriminative clustering. Experimental results demonstrate the superiority of our method. Future work will focus on extending this framework to accommodate incomplete multi-view scenarios and exploring accelerated sampling strategies to further enhance scalability.

Contribution Statement

*Jinli Ma and Chenkai Guo are co-first authors and contributed equally to this work, including conceiving the main idea, implementing the method, and conducting the experiments. Renda Han contributed to the framework illustration and algorithm development. Renxiang Guan, Siwei Wang, Ke Liang, and Xiaoyu Cui contributed to result analysis and manuscript revision. †Dayu Hu is the corresponding author and supervised the project. All authors reviewed and approved the final manuscript.

References

[Biroli *et al.*, 2024] Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

[Cai *et al.*, 2013] Xiao Cai, Feiping Nie, and Heng Huang. Multi-view k-means clustering on big data. In *IJCAI*, volume 13, pages 2598–2604, 2013.

[Chen *et al.*, 2022] Man-Sheng Chen, Jia-Qi Lin, Xiang-Long Li, Bao-Yu Liu, Chang-Dong Wang, Dong Huang, and Jian-Huang Lai. Representation learning in multi-view clustering: A literature review. *Data Science and Engineering*, 7(3):225–241, 2022.

[Croitoru *et al.*, 2023] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10850–10869, 2023.

[Ding *et al.*, 2025] Yu Ding, Katsuya Hotta, Chunzhi Gu, Ao Li, Jun Yu, and Chao Zhang. Learning to discriminate while contrasting: Combating false negative pairs with coupled contrastive learning for incomplete multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Fang *et al.*, 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.

[Fang *et al.*, 2025] Yuan Fang, Xiaofeng Feng, Geping Yang, Ruichu Cai, Yiyang Yang, Zhiguo Gong, and Zhifeng Hao. Eva-mvc: Equitable view-weight allocation for generic multi-view clustering. In *Proceedings of the ACM on Web Conference 2025*, pages 2992–3003, 2025.

[Fei *et al.*, 2025] Lunke Fei, Junlin He, Qi Zhu, Shuping Zhao, Jie Wen, and Yong Xu. Deep multi-view contrastive clustering via graph structure awareness. *IEEE Transactions on Image Processing*, 2025.

[Fu *et al.*, 2020] Lele Fu, Pengfei Lin, Athanasios V Vasilakos, and Shiping Wang. An overview of recent multi-view clustering. *Neurocomputing*, 402:148–161, 2020.

[Graham *et al.*, 2022] Matthew M Graham, Alexandre H Thiery, and Alexandros Beskos. Manifold markov chain monte carlo methods for bayesian inference in diffusion models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1229–1256, 2022.

[Gunawardena *et al.*, 2024] Sohan Gunawardena, Khanh Luong, Thirunavukarasu Balasubramaniam, and Richi Nayak. Dccnmf: Deep complementary and consensus non-negative matrix factorization for multi-view clustering. *Knowledge-Based Systems*, 285:111330, 2024.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

[Kang *et al.*, 2021] Zhao Kang, Zhiping Lin, Xiaofeng Zhu, and Wenbo Xu. Structured graph learning for scalable subspace clustering: From single view to multiview. *IEEE Transactions on Cybernetics*, 52(9):8976–8986, 2021.

[Kumar *et al.*, 2011] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. *Advances in neural information processing systems*, 24, 2011.

[Li *et al.*, 2018a] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, 19(2):325–340, 2018.

[Li *et al.*, 2018b] Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863–1883, 2018.

[Li *et al.*, 2023] Ang Li, Yawen Li, Yingxia Shao, and Bingyan Liu. Multi-view scholar clustering with dynamic interest tracking. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9671–9684, 2023.

[Li *et al.*, 2025] Zongwei Li, Lianghao Xia, Hua Hua, Shijie Zhang, Shuangyang Wang, and Chao Huang. Diffgraph: Heterogeneous graph diffusion model. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 40–49, 2025.

[Lu *et al.*, 2024] Han Lu, Huafu Xu, Qianqian Wang, Quanxue Gao, Ming Yang, and Xinbo Gao. Efficient multi-view -means for image clustering. *IEEE Transactions on Image Processing*, 33:273–284, 2024.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Shi *et al.*, 2025] Fuhao Shi, Shaohua Wan, Shengli Wu, Hui Wei, and Hu Lu. Deep contrastive coordinated multi-view consistency clustering. *Machine Learning*, 114(3):81, 2025.

[Smith and Dalchau, 2018] Stephen Smith and Neil Dalchau. Model reduction enables turing instability analysis of large reaction–diffusion models. *Journal of The Royal Society Interface*, 15(140):20170805, 2018.

[Song *et al.*, 2020] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[Sun *et al.*, 2024] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[Tang *et al.*, 2022] Kewei Tang, Kaiqiang Xu, Wei Jiang, Zhixun Su, Xiyan Sun, and Xiaonan Luo. Selecting the best part from multiple laplacian autoencoders for multi-view subspace clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7457–7469, 2022.

[Xi *et al.*, 2025] Yanwanyu Xi, Chang Tang, Jun-Jie Huang, Xingchen Hu, Yuanyuan Liu, and Xinwang Liu. Contrastive and dual adversarial representation learning for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Xu *et al.*, 2022] 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*, pages 16051–16060, 2022.

[Yan *et al.*, 2021] Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021.

[Yan *et al.*, 2023] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19863–19872, 2023.

[Yang *et al.*, 2023] Xihong Yang, Jin Jiaqi, Siwei Wang, Ke Liang, Yue Liu, Yi Wen, Suyuan Liu, Sihang Zhou, Xinwang Liu, and En Zhu. Dealmvc: Dual contrastive calibration for multi-view clustering. In *Proceedings of the 31st ACM international conference on multimedia*, pages 337–346, 2023.

[Yang *et al.*, 2025] Ben Yang, Xuetao Zhang, Jinghan Wu, Feiping Nie, Fei Wang, and Badong Chen. Scalable min-max multi-view spectral clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[Zhang *et al.*, 2024] Qian Zhang, Lin Zhang, Ran Song, Runmin Cong, Yonghuai Liu, and Wei Zhang. Learning common semantics via optimal transport for contrastive multi-view clustering. *IEEE Transactions on Image Processing*, 2024.

[Zhou *et al.*, 2024] Lihua Zhou, Guowang Du, Kevin Lue, Lizheng Wang, and Jingwei Du. A survey and an empirical evaluation of multi-view clustering approaches. *ACM Computing Surveys*, 56(7):1–38, 2024.