# DSMVC: DISENTANGLED STREAMING MULTI-VIEW CLUSTERING VIA PURIFICATION AND CALIBRATION

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

## ABSTRACT

Streaming multi-view clustering has attracted growing attention for its ability to aggregate continuously arriving views without accessing historical raw data. However, emerging views often introduce private variations and unreliable cross-view correspondences, inevitably corrupting the global consensus over time. To address this issue, we propose a novel Disentangled Streaming Multi-View Clustering (DSMVC) framework. Specifically, DSMVC leverages spectral curvature to decouple dominant structures from private residuals, securing reliable knowledge accumulation through two core modules: an Attribution-Guided Suppression Module that suppresses vulnerable components by quantifying feature attribution sensitivity to purify residual representations, and a Discrepancy-Guided Cross-view Alignment Module that calibrates cross-view correspondences via permutation-induced prediction inconsistency, ensuring robust consensus semantics across streaming views. Extensive experiments across benchmark datasets demonstrate that DSMVC achieves substantial advancement over state-of-the-art methods, exhibiting remarkable robustness against continuous view arrivals.

## 1 INTRODUCTION

Real-world entities are often characterized by heterogeneous features across diverse modalities, such as images, text descriptions, and audio signals (Liu et al., 2025; Xu & Wang, 2025; Wang et al., 2023). Multi-view clustering (MVC) aims to exploit the consistency and complementarity among these sources to partition samples into semantically distinct groups (Qin et al., 2025; Chowdhury et al., 2025; Li et al., 2025). Despite their progress, existing MVC paradigms predominantly assume full-view availability, ignoring the reality that data arrives sequentially over time. Taking medical diagnostics as an example, basic test results and CT scans can be obtained quickly, whereas genomic sequences may require days or weeks to become available. To address the challenge, streaming multi-view clustering (SMVC) methods have been proposed to incrementally accumulate cross-view knowledge without revisiting historical raw data (Wang et al., 2025; Zhang et al., 2024b).

Early approaches, such as IMSC (Zhou et al., 2019), integrate streaming views by learning a consensus kernel with low-rank approximation as views accumulate. Along this direction, CMVC (Wan et al., 2022) combines late fusion with continual learning and maintains a consensus partition matrix to incrementally fuse incoming views. Nevertheless, constructing full-scale affinity matrices incurs high computational costs, inevitably limiting their scalability on large datasets. To tackle this limitation, subsequent works introduce lightweight anchor structures and dynamic memory mechanisms. For instance, CMVC-PCAL (Wang & Lu, 2026) establishes an anchor-based alignment strategy that dynamically matches historical and current anchor graphs to preserve temporal consistency. Building on this, CKR-MVC (Liang et al., 2025) further incorporates contrastive knowledge replay to retain past semantics through cross-task distillation.

Despite these advances, existing SMVC frameworks implicitly assume that latent features are noise-free and cross-view associations strictly reflect semantic consensus. In practice, this premise rarely holds, exposing two critical challenges: **i) Representation Entanglement.** Due to the heterogeneity of data acquisition, incoming views naturally intertwine shared semantics with view-specific patterns and noise (Figure 1 (a)). Naively enforcing cross-view alignment inevitably encodes these private variations into the shared space, thereby corrupting the consensus representation. **ii) Unreliable Association.** Cross-view alignment often relies on shared yet non-discriminative context. As evident



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

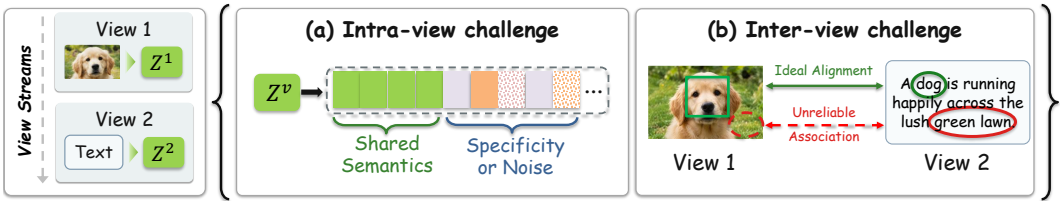



Figure 1: Motivation of DSMVC: streaming scenario with intra-view and inter-view challenges.

in Figure 1 (b), when a dog image and its text description repeatedly co-occur with a grassy field, models easily align the environmental background rather than the dog itself. Consequently, these entangled representations and spurious associations accumulate across streaming views, causing severe semantic degradation during continuous knowledge integration.

To overcome these challenges, we propose a novel Disentangled Streaming Multi-view Clustering (DSMVC) framework to secure reliable knowledge accumulation over continuous streams. Specifically, DSMVC first decouples each incoming view into dominant structures and private residuals via spectral curvature analysis. Subsequently, an Attribution-Guided Suppression Module (AGSM) is introduced to suppress vulnerable residual components by quantifying feature attribution sensitivity, preserving purified representations against noise interference. Furthermore, we devise a Discrepancy-Guided Cross-view Alignment Module (DGAM) to evaluate cross-view correspondence reliability. By contrasting naturally matched pairs against permuted pairs, DGAM dynamically regulates cross-view alignment via prediction inconsistency, eliminating spurious associations. Consequently, DSMVC effectively purifies streaming representations while calibrating cross-view consensus correspondences to achieve stable and discriminative clustering. Extensive experiments across benchmark datasets demonstrate the superiority and robustness of DSMVC over competitive state-of-the-art baselines. We summarize our main contributions as follows:

- We propose a novel Disentangled Streaming Multi-View Clustering that leverages spectral curvature analysis to decouple incoming views into dominant structures and private residuals, enabling reliable knowledge accumulation without revisiting historical raw data.
- We devise the Attribution-Guided Suppression Module to suppress vulnerable components via feature attribution sensitivity, yielding purified residual representations.
- We develop the Discrepancy-Guided Cross-view Alignment Module to dynamically calibrate cross-view correspondences. By contrasting naturally matched pairs against permuted pairs, this module measures prediction inconsistency to eliminate spurious associations and ensure robust consensus semantics.

## 2 Methodology

Given a multi-view dataset $\mathcal{X} = \{\mathbf{X}^v\}_{v=1}^V$ containing $N$ samples across $V$ views with dimensions $\{D^v\}_{v=1}^V$, streaming multi-view clustering aims to partition all instances into $K$ semantically coherent clusters without revisiting historical raw data. However, each emerging view often introduces private variations and unreliable cross-view correspondences, corrupting the global consensus over time. To tackle this, we propose the Disentangled Streaming Multi-view Clustering (DSMVC) framework. As illustrated in Figure 2, our DSMVC consists of three key modules: i) a disentangled feature learning module serving as our training baseline to separate dominant structures from private residuals, ii) an attribution-guided suppression module that purifies residual components to eliminate vulnerable factors, and iii) a discrepancy-guided cross-view alignment module that exploits permutation-induced prediction inconsistency to calibrate cross-view correspondences. By integrating these modules, DSMVC achieves reliable consensus accumulation and prevents detrimental perturbations, ensuring robust and discriminative cluster assignments.

### 2.1 Disentangled Feature Learning

In streaming scenarios, historical raw data cannot be revisited once new views arrive. Consequently, existing methods typically align latent representations to accumulate cross-view knowledge. However, real-world features inherently mix shared semantics with view-specific patterns and noise.

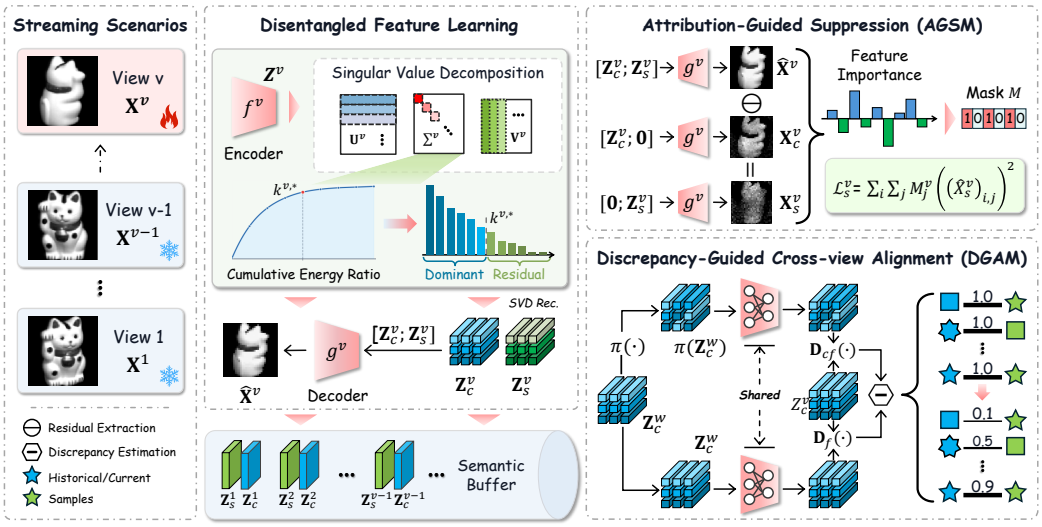



Figure 2: Overview of DSMVC.

Directly aligning these representations inevitably causes semantic distortion, disrupting established cluster structures. To tackle this, DSMVC explicitly separates the latent space of each arriving view into two components: a dominant representation $\mathbf{Z}_c^v \in \mathbb{R}^{N\times d}$ to capture the main structure of each incoming view, and a residual representation $\mathbf{Z}_s^v \in \mathbb{R}^{N\times d}$ that characterizes the remaining subtle variations. Specifically, when the $v$-th view $\mathbf{X}^v \in \mathbb{R}^{N\times D^v}$ arrives, an encoder network $f^v(\cdot;\theta^v)$ first projects the raw input into a unified latent space:

$$\mathbf{Z}^v = f^v(\mathbf{X}^v;\theta^v), \tag{1}$$

where $\mathbf{Z}^v \in \mathbb{R}^{N\times d}$, and $\theta^v$ denotes the learnable parameters of the encoder.

However, $\mathbf{Z}^v$ remains entangled with shared semantics, view-specific patterns, and noise, making direct decomposition difficult. Therefore, we introduce singular value decomposition to project this representation onto orthogonal geometric bases:

$$\mathbf{Z}^v = \mathbf{U}^v\mathbf{\Sigma}^v(\mathbf{V}^v)^\top = \sum_{i=1}^{r}\sigma_i^v\mathbf{u}_i^v(\mathbf{v}_i^v)^\top, \tag{2}$$

where $\mathbf{\Sigma}^v = \mathrm{diag}(\sigma_1^v,\ldots,\sigma_r^v)$ stores the singular values in descending order, with $r=\min(N,d)$.

Since singular values are sorted by energy in descending order, the top components naturally preserve the primary low-rank structure of $\mathbf{Z}^v$, whereas the trailing components retain subtle residual variations. To adaptively determine the optimal truncation boundary between them, we introduce a curvature-based spectral partition strategy via the cumulative energy ratio:

$$e_k^v = \frac{\sum_{i=1}^{k}(\sigma_i^v)^2}{\sum_{i=1}^{r}(\sigma_i^v)^2}, \qquad 1\le k\le r. \tag{3}$$

Following (Cattell, 1966), this cumulative ratio curve typically exhibits an elbow shape: it rises steeply at first and flattens after capturing the dominant variance. Thus, the optimal truncation boundary can be located at the point of maximum curvature. To elaborate, we map the discrete indices onto normalized coordinates $(x_i, y_i^v)$ for $i\in\{1,\ldots,r\}$, where $x_i=(i-1)/(r-1)\in[0,1]$ represents the normalized component index, and $y_i^v\in[0,1]$ denotes the rescaled cumulative energy:

$$y_i^v = \frac{e_i^v - e_1^v}{e_r^v - e_1^v}. \tag{4}$$

Subsequently, the discrete curvature $c_i^v$ is computed for each candidate point:

$$c_i^v = \frac{|\ddot{y}_i^v|}{(1+(\dot{y}_i^v)^2)^{3/2}}, \tag{5}$$

where $\dot{y}_i^v$ and $\ddot{y}_i^v$ denote the first and second numerical derivatives evaluated via central differences, respectively. Based on this, the optimal truncation boundary $k^{v,*}$ is determined at the peak curvature:

$$k^{v,*} = \arg\max_{1 \le i \le r} c_i^v. \tag{6}$$

This index $k^{v,*}$ marks the transition where marginal energy accumulation diminishes substantially, providing an intrinsic partition between principal patterns and subtle variations. Consequently, the latent representation is decoupled into:

$$\mathbf{Z}_c^v = \sum_{i=1}^{k^{v,*}} \sigma_i^v \mathbf{u}_i^v (\mathbf{v}_i^v)^\top, \qquad \mathbf{Z}_s^v = \sum_{i=k^{v,*}+1}^{r} \sigma_i^v \mathbf{u}_i^v (\mathbf{v}_i^v)^\top, \tag{7}$$

where $\mathbf{Z}_c^v, \mathbf{Z}_s^v \in \mathbb{R}^{N \times d}$ denote the dominant and residual representations, respectively. On this basis, we employ a decoder $g^v(\cdot; \phi^v)$ to reconstruct the original input from the concatenated representations:

$$\hat{\mathbf{X}}^v = g^v([\mathbf{Z}_c^v; \mathbf{Z}_s^v]; \phi^v), \tag{8}$$

where $[\cdot;\cdot]$ denotes column-wise feature concatenation, and $\phi^v$ encapsulates the trainable parameters of the decoder. To ensure the disentangled representations preserve sufficient input information, we minimize the reconstruction discrepancy on the incoming view:

$$\mathcal{L}_r^v = \frac{1}{N} \left\| \mathbf{X}^v - \hat{\mathbf{X}}^v \right\|_F^2. \tag{9}$$

In this way, each incoming representation is explicitly disentangled into a dominant component that captures the primary geometric structure, and a residual component that absorbs the remaining subtle variations. This complementary partitioning provides a well-defined foundation for subsequent cross-view alignment and noise purification.

### 2.2 Attribution-Guided Suppression Module

While disentangled representation learning separates the dominant components from residual variations, energy-based truncation solely discriminates variance magnitude rather than feature reliability. Consequently, the residual component inevitably contains vulnerable factors and noise dimensions that could distort clustering structures. To resolve this, we introduce the Attribution-Guided Suppression Module (AGSM) to identify and suppress detrimental factors. Instead of operating in the abstract latent space, AGSM assesses feature attribution within the reconstructed input space, providing an explicit physical anchor to quantify dimension-wise attribution. Specifically, we alternately zero-mask one branch to generate its individual reconstructions via the decoder:

$$\hat{\mathbf{X}}_c^v = g^v([\mathbf{Z}_c^v; \mathbf{0}]; \phi^v), \qquad \hat{\mathbf{X}}_s^v = g^v([\mathbf{0}; \mathbf{Z}_s^v]; \phi^v), \tag{10}$$

where $\hat{\mathbf{X}}_c^v, \hat{\mathbf{X}}_s^v \in \mathbb{R}^{N \times D^v}$ denote the reconstructed inputs generated solely from the dominant and residual components, respectively.

Ideally, under an additive decoding assumption, the difference $\mathbf{R}^v = \hat{\mathbf{X}}^v - \hat{\mathbf{X}}_c^v$ represents the complementary variation that should be captured by the residual reconstruction $\hat{\mathbf{X}}_s^v$. However, due to non-linear coupling within the decoder, $\hat{\mathbf{X}}_s^v$ inevitably deviates from $\mathbf{R}^v$. To quantify this reconstruction inconsistency, we formulate the discrepancy metric:

$$J^v = \frac{1}{N} \left\| \hat{\mathbf{X}}_s^v - \mathbf{R}^v \right\|_F^2. \tag{11}$$

Rather than directly minimizing $J^v$, we compute its first-order gradient attribution to trace dimension-wise contributions to this inconsistency. Concretely, the attribution vector $\mathbf{a}^v \in \mathbb{R}^{D^v}$ is derived by averaging the absolute gradient magnitudes across all samples:

$$\mathbf{a}^v = \frac{1}{N} \sum_{i=1}^{N} \left| (\nabla_{\hat{\mathbf{X}}_s^v} J^v)_i \right|, \tag{12}$$

where $a_j^v$ quantifies the attribution of the $j$-th feature dimension to reconstruction inconsistency. Intuitively, high-attribution dimensions dominate the residual inconsistency, reflecting unstable or

noisy variations. To isolate these detrimental factors, we compute the mean attribution threshold $\bar{a}^v = \frac{1}{D^v}\sum_{j=1}^{D^v} a_j^v$, and construct a binary suppression mask $M^v \in \{0,1\}^{D^v}$:

$$M_j^v = \begin{cases} 1, & \text{if } a_j^v > \bar{a}^v, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Here, $M_j^v = 1$ flags the $j$-th dimension as an over-contributing factor to the residual inconsistency, whereas dimensions with below-average attribution are preserved as informative view-private patterns. Accordingly, we penalize the energy of these volatile dimensions via the suppression loss:

$$\mathcal{L}_s^v = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{D^v} M_j^v \left((\hat{\mathbf{X}}_s^v)_{i,j}\right)^2. \tag{14}$$

Through backpropagation, the suppression loss propagates gradients back to the encoder via the decoder's Jacobian, where $\nabla_{\mathbf{Z}_s^v}\mathcal{L}_s^v = \nabla_{\hat{\mathbf{X}}_s^v}\mathcal{L}_s^v \frac{\partial \hat{\mathbf{X}}_s^v}{\partial \mathbf{Z}_s^v}$. By penalizing energy on the reconstructed dimensions flagged by $M_j^v = 1$, the encoder is guided to suppress the detrimental factors within $\mathbf{Z}_s^v$ that drive these conflicting patterns, effectively purifying the residual subspace.

### 2.3 Discrepancy-Guided Cross-view Alignment Module

The dominant representation reduces residual interference within each view, yet directly enforcing cross-view consistency still encounters unreliable associations. For example, observations across views frequently contain superficial correlations or ambiguous patterns, which mislead naive alignment into capturing spurious shortcuts rather than intrinsic cluster semantics. To address this, we propose the Discrepancy-Guided Cross-View Alignment Module (DGAM). By comparing naturally matched pairs with randomly permuted counterparts, this module estimates the reliability of each cross-view correspondence to ensure intrinsic consensus in streaming scenarios. Specifically, let $\mathbf{Z}_c^w, \mathbf{Z}_c^v \in \mathbb{R}^{N\times d}$ denote the dominant representations from a historical view $w$ and the current view $v$, respectively. We introduce a parametric predictor $\mathcal{F}^v(\cdot)$ to project $\mathbf{Z}_c^w$ into the semantic space of view $v$, thereby calculating the predictive discrepancy between naturally paired representations:

$$(\mathbf{D}_f^w)_i = \frac{1}{d}\left\|\mathcal{F}^v(\mathbf{Z}_c^w)_i - (\mathbf{Z}_c^v)_i\right\|_2^2, \tag{15}$$

where $(\mathbf{Z}_c^v)_i \in \mathbb{R}^d$ denotes the dominant representation of the $i$-th instance in view $v$, and $\mathcal{F}^v(\mathbf{Z}_c^w)_i \in \mathbb{R}^d$ represents its corresponding projection from view $w$. Here, $\mathbf{D}_f^w \in \mathbb{R}^N$ collects the alignment discrepancy across all natural pairs.

To evaluate the semantic reliability of this cross-view connection, we construct an unaligned scenario by shuffling the natural correspondence. In detail, we apply a random permutation $\pi(\cdot)$ over the sample indices such that $\pi(i) \neq i$. The resulting permuted-pair prediction discrepancy is defined as:

$$(\mathbf{D}_{cf}^w)_i = \frac{1}{d}\left\|\mathcal{F}^v((\mathbf{Z}_c^w)_{\pi(i)}) - (\mathbf{Z}_c^v)_i\right\|_2^2, \tag{16}$$

where $(\mathbf{Z}_c^w)_{\pi(i)} \in \mathbb{R}^d$ represents the mismatched feature from view $w$.

The underlying rationale is straightforward: for reliably associated instances, the aligned discrepancy $(\mathbf{D}_f^w)_i$ is small, while shuffling disrupts the pairing and causes a higher discrepancy $(\mathbf{D}_{cf}^w)_i > (\mathbf{D}_f^w)_i$. Conversely, for ambiguous or weakly correlated pairs, shuffling leads to negligible variation, yielding $(\mathbf{D}_{cf}^w)_i \approx (\mathbf{D}_f^w)_i$. Accordingly, the gap between the two discrepancies reflects the correspondence reliability of the $i$-th instance:

$$\Delta_i^w = (\mathbf{D}_{cf}^w)_i - (\mathbf{D}_f^w)_i. \tag{17}$$

Next, we compute an adaptive weight via $\gamma_i^w = \mathrm{Sigmoid}(\Delta_i^w)$, which assigns higher weights to reliable correspondences while down-weighting uncertain ones. Building on this, the discrepancy-guided alignment loss across all $W$ historical views and the current view $v$ is formulated as:

$$\mathcal{L}_c^v = \frac{1}{W}\sum_{w=1}^{W}\frac{1}{N}\sum_{i=1}^{N}\gamma_i^w(\mathbf{D}_f^w)_i. \tag{18}$$

Through this discrepancy-guided calibration, DGAM adaptively prioritizes reliable cross-view correspondences and eliminates spurious associations, ultimately ensuring robust consensus semantics across streaming views.

### 2.4 Overall Objective and Training Strategy

With the above components, the overall objective function for training the incoming view $v$ is formulated as:

$$\mathcal{L}^v = \mathcal{L}_r^v + \lambda_1 \mathcal{L}_s^v + \lambda_2 \mathcal{L}_c^v, \tag{19}$$

where $\lambda_1$ and $\lambda_2$ are trade-off hyperparameters. Notably, in the streaming setting, the initial view ($v = 1$) is optimized with $\lambda_2 = 0$. For subsequent views ($v > 1$), all three terms are jointly minimized to achieve reconstruction, noise suppression, and reliability-aware cross-view alignment.

Table 1: Clustering results on seven benchmark datasets. Among *Sequence Access* methods, the best and second-best results are highlighted in **bold** and underline, respectively. Among *Full Access* methods, the best result is marked by $*$.

| Methods | COIL20 | ALOI | 100Leaves | Handwritten | Out-Scene | ALOI100 | Cifar100 |
|---|---|---|---|---|---|---|---|
| **ACC (%)** | | | | | | | |
| *Full Access Methods* | | | | | | | |
| SCMVC (TMM'24) | $77.79 \pm 5.10$ | $89.69 \pm 2.68$ | $83.19 \pm 1.10$ | $92.69 \pm 1.36$ | $73.72 \pm 0.61^*$ | $87.85 \pm 0.38^*$ | $98.77 \pm 0.55$ |
| MGCMVC (TNNLS'25) | $92.87 \pm 1.60$ | $80.24 \pm 1.50$ | $47.73 \pm 0.58$ | $88.19 \pm 5.93$ | $66.37 \pm 1.49$ | $62.08 \pm 1.52$ | $68.26 \pm 1.24$ |
| DFL-Net (TKDE'25) | $22.00 \pm 2.80$ | $37.28 \pm 4.03$ | $17.35 \pm 1.81$ | $64.08 \pm 5.81$ | $53.83 \pm 2.32$ | $29.02 \pm 3.19$ | $33.01 \pm 2.00$ |
| BONE (AAAI'26) | $96.41 \pm 2.33^*$ | $88.50 \pm 3.33$ | $82.18 \pm 1.05$ | $91.83 \pm 4.51$ | $70.78 \pm 6.05$ | $80.69 \pm 2.30$ | $99.30 \pm 0.48$ |
| WAH (AAAI'26) | $86.96 \pm 3.57$ | $88.34 \pm 2.09$ | $35.62 \pm 2.71$ | $82.95 \pm 4.65$ | $65.18 \pm 2.70$ | $62.21 \pm 1.60$ | $82.79 \pm 2.50$ |
| DMVCS (TKDE'26) | $82.79 \pm 2.52$ | $96.33 \pm 0.73^*$ | $92.53 \pm 1.30^*$ | $98.53 \pm 0.25^*$ | $70.84 \pm 0.98$ | $85.64 \pm 1.79$ | $99.38 \pm 0.79^*$ |
| *Sequence Access Methods* | | | | | | | |
| LAIMVC (MM'24) | $61.56 \pm 3.67$ | $85.01 \pm 5.74$ | $50.91 \pm 1.35$ | $61.27 \pm 4.93$ | $28.15 \pm 2.47$ | $44.79 \pm 0.79$ | $84.89 \pm 1.44$ |
| FDMVC (KBS'24) | $77.79 \pm 3.64$ | $76.73 \pm 1.06$ | $76.86 \pm 1.53$ | $87.90 \pm 1.90$ | $66.66 \pm 0.41$ | $62.22 \pm 1.47$ | $84.54 \pm 1.36$ |
| DSVC (ICML'25) | $73.66 \pm 2.54$ | $89.65 \pm 5.49$ | $87.05 \pm 1.95$ | $86.90 \pm 6.41$ | $58.95 \pm 6.62$ | $81.40 \pm 0.83$ | $89.30 \pm 1.43$ |
| ADIMC (NN'26) | $73.54 \pm 1.94$ | $71.18 \pm 0.00$ | $75.51 \pm 2.16$ | $76.07 \pm 0.30$ | $68.36 \pm 0.98$ | $50.75 \pm 1.34$ | $88.95 \pm 1.54$ |
| CMSA (PR'26) | $59.08 \pm 2.72$ | $69.40 \pm 4.48$ | $42.81 \pm 2.48$ | $45.76 \pm 1.75$ | $39.49 \pm 11.73$ | $47.84 \pm 1.96$ | $95.77 \pm 1.11$ |
| DSMVC (Ours) | $\mathbf{79.71 \pm 1.76}$ | $\mathbf{94.53 \pm 0.56}$ | $\mathbf{90.41 \pm 1.38}$ | $\mathbf{96.67 \pm 0.36}$ | $\mathbf{69.63 \pm 4.85}$ | $\mathbf{84.76 \pm 0.94}$ | $\mathbf{96.35 \pm 0.68}$ |
| **NMI (%)** | | | | | | | |
| *Full Access Methods* | | | | | | | |
| SCMVC (TMM'24) | $87.73 \pm 2.49$ | $86.50 \pm 2.71$ | $92.40 \pm 0.58$ | $86.62 \pm 1.81$ | $61.07 \pm 0.26$ | $92.74 \pm 0.31$ | $99.71 \pm 0.08$ |
| MGCMVC (TNNLS'25) | $94.54 \pm 1.09$ | $77.47 \pm 1.37$ | $75.73 \pm 0.31$ | $86.44 \pm 3.86$ | $55.55 \pm 0.62$ | $82.24 \pm 0.77$ | $89.29 \pm 0.12$ |
| DFL-Net (TKDE'25) | $48.36 \pm 2.88$ | $48.67 \pm 2.83$ | $51.33 \pm 2.81$ | $61.44 \pm 3.59$ | $50.09 \pm 1.42$ | $62.04 \pm 2.06$ | $72.00 \pm 1.87$ |
| BONE (AAAI'26) | $98.30 \pm 0.71^*$ | $86.86 \pm 2.17$ | $90.48 \pm 0.68$ | $90.41 \pm 2.68$ | $59.62 \pm 3.55$ | $86.62 \pm 0.85$ | $99.57 \pm 0.14$ |
| WAH (AAAI'26) | $93.65 \pm 2.13$ | $85.31 \pm 2.75$ | $68.03 \pm 1.89$ | $86.15 \pm 1.66$ | $54.88 \pm 1.21$ | $83.63 \pm 0.60$ | $90.19 \pm 0.94$ |
| DMVCS (TKDE'26) | $90.20 \pm 1.02$ | $93.51 \pm 0.91^*$ | $97.19 \pm 0.34^*$ | $96.46 \pm 0.45^*$ | $61.34 \pm 1.23^*$ | $92.77 \pm 0.40^*$ | $99.85 \pm 0.12^*$ |
| *Sequence Access Methods* | | | | | | | |
| LAIMVC (MM'24) | $73.07 \pm 1.60$ | $83.09 \pm 3.17$ | $71.80 \pm 0.99$ | $58.62 \pm 1.62$ | $14.65 \pm 2.42$ | $66.24 \pm 0.36$ | $95.80 \pm 0.51$ |
| FDMVC (KBS'24) | $85.25 \pm 1.23$ | $76.85 \pm 0.53$ | $90.07 \pm 0.49$ | $80.52 \pm 1.76$ | $50.99 \pm 1.85$ | $76.36 \pm 0.34$ | $96.41 \pm 0.51$ |
| DSVC (ICML'25) | $83.26 \pm 0.83$ | $87.50 \pm 2.47$ | $94.98 \pm 0.71$ | $85.77 \pm 2.94$ | $50.58 \pm 3.81$ | $88.90 \pm 0.11$ | $96.44 \pm 0.35$ |
| ADIMC (NN'26) | $80.09 \pm 1.24$ | $66.44 \pm 1.17$ | $89.48 \pm 0.57$ | $66.81 \pm 0.16$ | $51.82 \pm 1.19$ | $68.48 \pm 0.57$ | $97.52 \pm 0.37$ |
| CMSA (PR'26) | $70.94 \pm 0.91$ | $68.64 \pm 3.16$ | $69.75 \pm 1.28$ | $40.47 \pm 2.20$ | $33.64 \pm 14.49$ | $73.82 \pm 1.65$ | $98.08 \pm 0.84$ |
| DSMVC (Ours) | $\mathbf{86.87 \pm 1.53}$ | $\mathbf{90.39 \pm 1.02}$ | $\mathbf{96.16 \pm 0.58}$ | $\mathbf{92.55 \pm 0.71}$ | $\mathbf{55.65 \pm 2.75}$ | $\mathbf{90.26 \pm 0.42}$ | $\mathbf{99.24 \pm 0.12}$ |
| **ARI (%)** | | | | | | | |
| *Full Access Methods* | | | | | | | |
| SCMVC (TMM'24) | $72.29 \pm 5.15$ | $80.89 \pm 4.46$ | $76.37 \pm 1.48$ | $84.85 \pm 2.61$ | $52.47 \pm 0.54^*$ | $82.23 \pm 0.71^*$ | $98.82 \pm 0.47$ |
| MGCMVC (TNNLS'25) | $88.45 \pm 1.83$ | $67.94 \pm 2.04$ | $36.05 \pm 1.05$ | $82.02 \pm 6.59$ | $43.92 \pm 1.38$ | $55.77 \pm 1.51$ | $61.77 \pm 1.08$ |
| DFL-Net (TKDE'25) | $15.49 \pm 2.72$ | $24.79 \pm 2.89$ | $8.27 \pm 1.70$ | $49.10 \pm 4.61$ | $36.82 \pm 2.08$ | $20.36 \pm 3.36$ | $26.93 \pm 3.66$ |
| BONE (AAAI'26) | $95.41 \pm 2.51^*$ | $80.80 \pm 3.65$ | $74.29 \pm 1.74$ | $88.10 \pm 5.01$ | $50.91 \pm 5.15$ | $71.51 \pm 2.29$ | $99.01 \pm 0.53$ |
| WAH (AAAI'26) | $82.92 \pm 5.23$ | $79.30 \pm 3.67$ | $23.93 \pm 2.27$ | $78.32 \pm 4.25$ | $43.43 \pm 2.21$ | $51.88 \pm 2.80$ | $78.94 \pm 1.92$ |
| DMVCS (TKDE'26) | $78.94 \pm 2.01$ | $91.89 \pm 1.66^*$ | $90.06 \pm 1.27^*$ | $96.76 \pm 0.54^*$ | $50.86 \pm 1.48$ | $81.64 \pm 1.43$ | $99.42 \pm 0.68^*$ |
| *Sequence Access Methods* | | | | | | | |
| LAIMVC (MM'24) | $52.58 \pm 3.04$ | $77.86 \pm 5.60$ | $33.93 \pm 1.83$ | $44.18 \pm 3.46$ | $6.92 \pm 1.20$ | $30.14 \pm 0.72$ | $85.98 \pm 1.39$ |
| FDMVC (KBS'24) | $72.19 \pm 3.06$ | $64.80 \pm 0.46$ | $70.47 \pm 1.57$ | $76.27 \pm 2.32$ | $43.69 \pm 2.63$ | $50.41 \pm 1.08$ | $86.58 \pm 1.34$ |
| DSVC (ICML'25) | $68.06 \pm 2.18$ | $83.61 \pm 4.66$ | $83.05 \pm 2.42$ | $81.50 \pm 5.89$ | $41.16 \pm 4.49$ | $74.63 \pm 0.16$ | $89.07 \pm 1.17$ |
| ADIMC (NN'26) | $65.96 \pm 1.87$ | $56.91 \pm 2.91$ | $68.77 \pm 1.85$ | $57.01 \pm 0.33$ | $45.20 \pm 1.35$ | $37.36 \pm 1.39$ | $90.18 \pm 1.39$ |
| CMSA (PR'26) | $46.74 \pm 2.31$ | $55.78 \pm 4.72$ | $28.73 \pm 2.18$ | $27.40 \pm 2.61$ | $21.54 \pm 11.76$ | $36.19 \pm 2.12$ | $95.09 \pm 1.44$ |
| DSMVC (Ours) | $\mathbf{72.82 \pm 2.34}$ | $\mathbf{87.81 \pm 1.32}$ | $\mathbf{87.19 \pm 1.69}$ | $\mathbf{92.76 \pm 0.77}$ | $\mathbf{47.74 \pm 3.67}$ | $\mathbf{77.45 \pm 1.21}$ | $\mathbf{96.67 \pm 0.56}$ |

## 3 Experiments

### 3.1 Experimental Settings

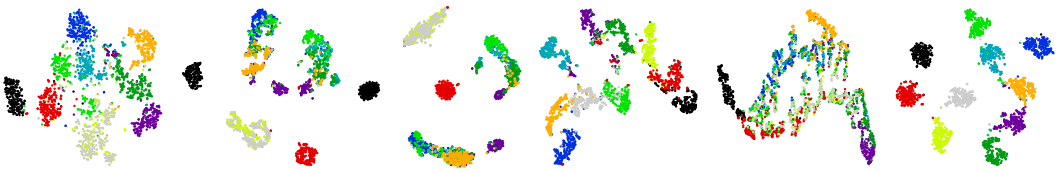
We evaluate DSMVC on seven multi-view benchmarks, namely COIL20 (Nene et al., 1996), ALOI (Geusebroek et al., 2005), 100Leaves (Mallah et al., 2013), Handwritten (Breukelen et al., 1998), Out-Scene (Oliva & Torralba, 2001), ALOI100 (Geusebroek et al., 2005), and Cifar100 (Wu et al., 2024). Following standard evaluation protocols, we adopt Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) as performance metrics, where higher values indicate superior clustering quality. To ensure a comprehensive comparison, we benchmark

DSMVC against eleven competitive methods under two paradigms: **i) Full Access Methods**, which require simultaneous availability of all views, including SCMVC (Wu et al., 2024), MGCMVC (Ju et al., 2025), DFL-Net (Chen et al., 2025), BONE (Xu et al., 2026), WAH (Wang et al., 2026), and DMVCS (Li et al., 2026). **ii) Sequence Access Methods**, which process streaming views sequentially, comprising LAIMVC (Qu et al., 2024), FDMVC (Lu et al., 2024), DSVC (Yuan et al., 2025), ADIMC (Zhong et al., 2026), and CMSA (Yin et al., 2026).

### 3.2 Comparative Results Analysis

Table 1 reports the clustering performance of all comparative methods, where each result is presented as the mean $\pm$ standard deviation over five independent runs. From the empirical results, we draw the following key observations: i) **Superiority in the Sequence-Access Setting.** DSMVC consistently achieves the best performance across all seven benchmarks and three evaluation metrics among sequence-access methods. Notably, on the Handwritten dataset, DSMVC surpasses the second-best sequential baseline by substantial margins of 8.77% in ACC, 6.78% in NMI, and 11.26% in ARI. These gains demonstrate that DSMVC effectively learns discriminative representations by purifying sequentially arriving views and calibrating cross-view consensus correspondences. ii) **Competitiveness against Full-Access Methods.** Although full-access methods can simultaneously exploit all views during training, DSMVC exhibits comparable or even superior performance under the more restrictive streaming setting. For instance, DSMVC achieves 94.53% ACC on ALOI and 96.67% ACC on Handwritten, outperforming the majority of full-access algorithms. This confirms that integrating intra-view noise suppression with reliability-aware alignment can preserve and accumulate cross-view semantics, substantially mitigating error propagation and catastrophic forgetting.

Furthermore, we conduct t-SNE visualizations on the Handwritten dataset to examine the embedding structure of the learned representations. As illustrated in Figure 3, the raw features present ambiguous cluster boundaries with severe category overlap. While streaming baselines partially alleviate this issue, recent methods like ADIMC and DSVC still exhibit obvious local confusion. In contrast, DSMVC demonstrates tighter intra-class compactness and clearer inter-class separation, confirming that our framework can reliably preserve structural discriminability as views arrive sequentially.



(a) Raw (b) LAIMVC (c) DSVC (d) ADIMC (e) CMSA (f) DSMVC

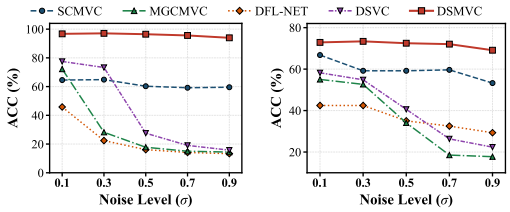
Figure 3: t-SNE visualizations of final representations on Handwritten across various methods.

### 3.3 Robustness Analysis to Gaussian Noise

We examine the robustness of DSMVC against input corruptions by injecting Gaussian noise with varying standard deviations $\sigma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ into each view of the Handwritten and Out-Scene datasets. As illustrated in Figure 4, DSMVC consistently achieves superior clustering performance across the entire noise range on both benchmarks. More importantly, while the comparative methods suffer from substantial performance degradation as the noise level increases, DSMVC exhibits only a slight decline, demonstrating stronger robustness under severe corruption. This advantage mainly stems from two coordinated mechanisms: i) AGSM purifies residual representations by suppressing vulnerable components via feature attribution sensitivity, and ii) DGAM calibrates cross-view correspondences using permutation-induced prediction inconsistency to eliminate unreliable associations. Their combination effectively prevents error accumulation across streaming views, enabling DSMVC to maintain robust consensus semantics even under strong perturbations.

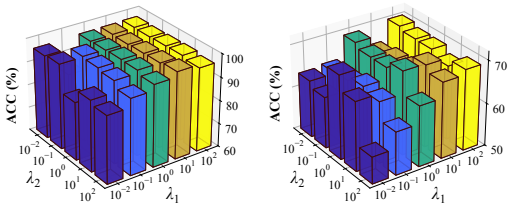
### 3.4 Hyper-parameter Analysis

As shown in Figure 5, we investigate the sensitivity of the trade-off hyperparameters $\lambda_1$ and $\lambda_2$ in Eq. (19). Specifically, on the Handwritten dataset, the clustering ACC remains relatively stable




(a) Handwritten (b) Out-Scene

Figure 4: Clustering performance under different levels of Gaussian noise.




(a) Handwritten (b) Out-Scene

Figure 5: Sensitivity analysis of $\lambda_1$ and $\lambda_2$ on two benchmark datasets.

across a broad range of values from $10^{-2}$ to $10^{2}$. Conversely, the results on the Out-Scene dataset exhibit significant fluctuations under varying combinations of $\lambda_1$ and $\lambda_2$. This occurs because an excessively large $\lambda_1$ over-penalizes discriminative features, whereas an undersized $\lambda_2$ hinders knowledge transfer across streaming views, thus failing to preserve shared semantics. Based on these, we empirically adopt $\lambda_1 = 1.0$ and $\lambda_2 = 0.1$ as the default configuration for all benchmarks.

### 3.5 Qualitative Analysis of Representation Disentanglement

As shown in Figure 6, we examine the qualitative behavior of our representation decomposition on the COIL20 dataset. While raw observations (first column) inherently interweave core semantic structures with view-specific visual variations, our framework explicitly separates the dominant representations from residual counterparts. Specifically, the dominant reconstructions (second column) reliably capture global object geometry and key contours, preserving essential category semantics. In contrast, the residual components (third column) collect high-frequency fluctuations and diffuse illumination speckles, suggesting that view-private perturbations are effectively concentrated in the residual branch. Furthermore, by adaptively suppressing volatile dimensions via AGSM, the purified reconstructions (fourth column) substantially eliminate background interference while retaining sharp object contours. These results confirm that DSMVC effectively disentangles dominant semantic structures from view-private variations, with AGSM further filtering residual instability.

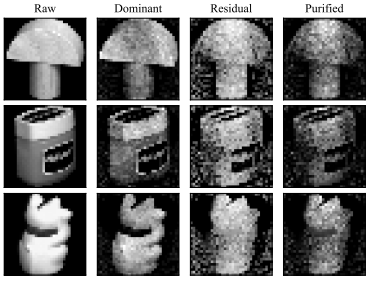
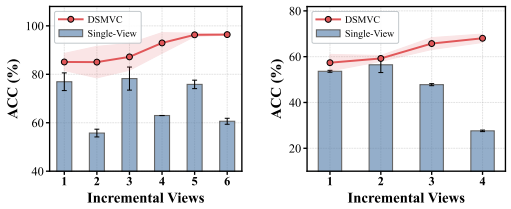



Figure 6: Visual disentanglement on COIL20: raw, dominant, residual, and purified features.

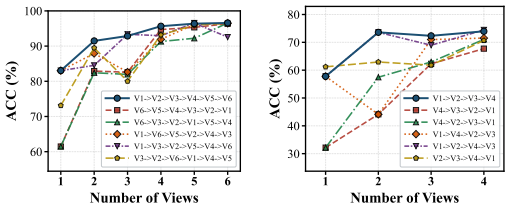
### 3.6 Stability Analysis under Incremental Views

We evaluate the anti-forgetting capability and sequential knowledge accumulation of DSMVC on the Handwritten and Out-Scene datasets. As depicted in Figure 7, the clustering accuracy of DSMVC improves steadily as more views arrive, and the fused representations consistently outperform individual single-view counterparts throughout the streaming process. Notably, although the quality of individual views fluctuates, DSMVC remains stable or even advances when incorporating low-quality views, confirming that our framework effectively resists catastrophic forgetting. This resilience is driven by three collaborative modules. First, the semantic buffer preserves compact historical representations to serve as a reliable reference for sequential updates. Second, AGSM actively purifies incoming residual representations by suppressing vulnerable feature dimensions based on attribution sensitivity. Finally, DGAM evaluates the discrepancy between matched and permuted pairs to modulate cross-view correspondences, ensuring faithful knowledge transfer across views. By integrating these modules, DSMVC successfully accumulates complementary semantics from streaming views while preventing catastrophic forgetting and individual view degradation.




(a) Handwritten (b) Out-Scene

Figure 7: Performance comparison between incremental views and single views.




(a) Handwritten (b) Out-Scene

Figure 8: Clustering sensitivity evaluation under varying view arrival sequences.

### 3.7 Sensitivity Analysis to View Arrival Orders

To investigate the sensitivity of DSMVC to the view arrival sequence, we evaluate different view orders on the Handwritten and Out-Scene datasets. As shown in Figure 8, DSMVC exhibits remarkable robustness to varying arrival orders, consistently converging to competitive and stable clustering performance across all sequences. Although minor performance variations occur in the earliest streaming stages due to differences in initial view quality, this gap diminishes rapidly as subsequent views are absorbed. This order-agnostic property primarily benefits from the semantic buffer and our DGAM, which preserves historical representations and calibrates cross-view correspondence to provide a stable semantic anchor for newly arriving views. Consequently, DSMVC is highly practical for real-world streaming scenarios where view sequences are inherently unpredictable.

### 3.8 Ablation Study

To evaluate the individual contributions of AGSM and DGAM, we conduct ablation studies on the 100Leaves, Handwritten, and Out-Scene datasets by removing their corresponding loss objectives $\mathcal{L}_s$ and $\mathcal{L}_c$, respectively. As reported in Table 2, the complete model DSMVC consistently achieves the highest clustering performance across all evaluation metrics, whereas removing either objective leads to clear performance drops. Specifically, the variant without $\mathcal{L}_s$ incurs severe performance degradation, particularly on Out-Scene, where the clustering ACC decreases sharply by 11.05% and ARI drops by 10.27%. This substantial decline confirms that AGSM plays a critical role in purifying residual representations by suppressing vulnerable feature components. Meanwhile, excluding $\mathcal{L}_c$ consistently degrades clustering accuracy across all benchmarks (e.g., dropping ACC by 2.20% on Out-Scene and 1.72% on Handwritten), indicating that DGAM effectively calibrates cross-view correspondences to mitigate unreliable associations and enhance semantic consensus. Consequently, both modules are necessary and complementary, collaboratively ensuring the effectiveness of DSMVC in streaming scenarios.

Table 2: Ablation study on key components of DSMVC.

| Variant | 100Leaves | | | Handwritten | | | Out-Scene | | |
|---|---|---|---|---|---|---|---|---|---|
| | ACC | NMI | ARI | ACC | NMI | ARI | ACC | NMI | ARI |
| w/o $\mathcal{L}_s$ | 88.44 | 95.85 | 83.83 | 95.35 | 90.98 | 89.90 | 58.58 | 46.41 | 37.47 |
| w/o $\mathcal{L}_c$ | 89.69 | 95.69 | 85.77 | 94.95 | 90.02 | 89.07 | 67.43 | 53.19 | 45.38 |
| DSMVC | **90.41** | **96.16** | **87.19** | **96.67** | **92.55** | **92.76** | **69.63** | **55.65** | **47.74** |

## 4 Conclusion

In this paper, we propose DSMVC, a novel disentangled streaming multi-view clustering framework to secure reliable knowledge accumulation across sequential views without accessing historical raw data. Specifically, DSMVC leverages spectral curvature to decouple incoming views into dominant structures and private residuals. Building on this, the Attribution-Guided Suppression Module purifies residual representations by suppressing vulnerable components via feature attribution sensitivity, preserving fine-grained variants against noise interference. Furthermore, the Discrepancy-Guided Cross-view Alignment Module dynamically calibrates cross-view correspondences using permutation-induced prediction inconsistency to eliminate unreliable associations, ensuring intrinsic consensus across streaming views. Comprehensive experiments across multiple benchmark datasets demonstrate that DSMVC consistently outperforms state-of-the-art baselines, thereby achieving remarkable robustness and stability against continuous view arrivals.

## AI USE STATEMENT

In this work, we used generative AI tools for language editing and text polishing. We have not used generative AI tools for research ideation or method design, and data collection are not applicable to this work. Additionally, we used generative AI tools for improving the overall readability and flow of the manuscript. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

## REFERENCES

M. V. Breukelen, R. Duin, D. Tax, and J. D. Hartog. Handwritten digit recognition by combined classifiers. *Kybernetika*, 1998.

Raymond B Cattell. The scree test for the number of factors. *Multivariate behavioral research*, 1 (2):245–276, 1966.

Jingyi Chen, Zhibin Dong, Tiejun Li, and Yibo Han. Dc-span: A dual contrastive attention network for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20127–20135, 2026.

Zhe Chen, Xiao-Jun Wu, Tianyang Xu, and Josef Kittler. Dfl-net: Disentangled feature learning network for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 37 (8):4537–4547, 2025.

Anal Roy Chowdhury, Avisek Gupta, and Swagatam Das. Deep multi-view clustering: A comprehensive survey of the contemporary techniques. *Information Fusion*, 119:103012, 2025.

Yu Feng, Weixuan Liang, Xinhang Wan, Jiyuan Liu, Miaomiao Li, and Xinwang Liu. Incremental multi-view clustering: Exploring stream-view correlations to learn consistency and diversity. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

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

Hengrong Ju, Yang Lu, Weiping Ding, Wei Zhang, and Xibei Yang. Multigranularity information fused contrastive learning with multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Guanzhou Ke, Yang Yu, Guoqing Chao, Xiaoli Wang, Chenyang Xu, and Shengfeng He. Disentangling multi-view representations beyond inductive bias. In *Proceedings of the ACM International Conference on Multimedia*, pp. 2582–2590, 2023.

Guanzhou Ke, Bo Wang, Xiaoli Wang, and Shengfeng He. Rethinking multi-view representation learning via distilled disentangling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26764–26773, 2024.

Haoran Li, Zhenwen Ren, Yulan Guo, Jiali You, and Xiaojian You. Lsvc: A lifelong learning approach for stream-view clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5):8714–8727, 2024.

Pengyuan Li, Man Liu, Dongxia Chang, Yiming Wang, Zisen Kong, and Yao Zhao. Aemvc: Mitigate imbalanced embedding space in multi-view clustering. In *Proceedings of the ACM International Conference on Multimedia*, pp. 6461–6470, 2025.

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

Haoyang Liang, Bingjun Wei, Xuemei Cao, Jiafen Liu, Xiaocao Ouyang, Jiangtao Qiu, Hao Wang, Xin Yang, and Tianrui Li. Continual deep multi-view clustering via contrastive knowledge replay. *Pattern Recognition*, 172:112548, 2025.

Xinwang Liu, Ke Liang, Jun Wang, Suyuan Liu, Xiangke Wang, and Huaimin Wang. Two decades of multi-view clustering: Taxonomy, application, and challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2025.

Shuyao Lu, Deng Xu, Chao Zhang, and Zhangqing Zhu. Fast dynamic multi-view clustering with semantic-consistency inheritance. *Knowledge-Based Systems*, 300:112247, 2024.

Charles Mallah, James Cope, James Orwell, et al. Plant leaf classification using probabilistic integration of shape, texture and margin features. *Signal Processing, Pattern Recognition and Applications*, 5(1):45–54, 2013.

SA Nene, SK Nayar, and H Murase. Columbia object image library (coil-20). Technical report, CUCS-005-96, February, 1996.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

Yalan Qin, Xinpeng Zhang, Shui Yu, and Guorui Feng. A survey on representation learning for multi-view data. *Neural Networks*, 181:106842, 2025.

Qian Qu, Xinhang Wan, Weixuan Liang, Jiyuan Liu, Yu Feng, Huiying Xu, Xinwang Liu, and En Zhu. A lightweight anchor-based incremental framework for multi-view clustering. In *Proceedings of the ACM International Conference on Multimedia*, pp. 8652–8661, 2024.

Gan Sun, Yang Cong, Yulun Zhang, Guoshuai Zhao, and Yun Fu. Continual multiview task learning via deep matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems*, 32 (1):139–150, 2020.

Xinhang Wan, Jiyuan Liu, Weixuan Liang, Xinwang Liu, Yi Wen, and En Zhu. Continual multi-view clustering. In *Proceedings of the ACM International Conference on Multimedia*, pp. 3676–3684, 2022.

Fangdi Wang, Jiaqi Jin, Zhibin Dong, Xihong Yang, Yu Feng, Xinwang Liu, Xinzhong Zhu, Siwei Wang, Tianrui Liu, and En Zhu. View gap matters: Cross-view topology and information decoupling for multi-view clustering. In *Proceedings of the ACM International Conference on Multimedia*, pp. 8431–8440, 2024.

Jing Wang, Songhe Feng, Kristoffer Knutsen Wickstrøm, and Michael C Kampffmeyer. Adaptcmvc: Robust adaption to incremental views in continual multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10285–10294, 2025.

Maoyuan Wang and Gui-Fu Lu. Continual multi-view clustering with pseudo-label guided consistency anchor learning. *Neural Networks*, pp. 109385, 2026.

Rui Wang, Yuting Jiang, Xiaoqing Luo, Xiao-Jun Wu, Nicu Sebe, and Ziheng Chen. Wasserstein-aligned hyperbolic multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26444–26452, 2026.

Yiming Wang, Dongxia Chang, Zhiqiang Fu, and Yao Zhao. Consistent multiple graph embedding for multi-view clustering. *IEEE Transactions on Multimedia*, 25:1008–1018, 2023.

Song Wu, Yan Zheng, Yazhou Ren, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, and Lifang He. Self-weighted contrastive fusion for deep multi-view clustering. *IEEE Transactions on Multimedia*, pp. 1–13, 2024.

Deng Xu, Chao Zhang, Zechao Li, Chunlin Chen, and Huaxiong Li. Fast disentangled slim tensor learning for multi-view clustering. *IEEE Transactions on Multimedia*, 27:1254–1265, 2024.

HaiMing Xu and Qianqian Wang. Llm-damvc: A large language model assisted dynamic agent for multi-view clustering. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2025.

Huilang Xu, Xiang-Xiang Su, Simin Chen, Guang-Yong Chen, and Xing Chen. Bridging optimization and neural networks for efficient multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16066–16074, 2026.

Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9214–9223, 2021.

Xiaoqiang Yan, Zhixiang Jin, Fengshou Han, and Yangdong Ye. Differentiable information bottleneck for deterministic multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27435–27444, 2024.

Hongwei Yin, Hongpeng Chen, Wenjun Hu, Peng Zhou, and Shitong Wang. Contrast-driven incremental multi-view clustering with semantic distillation and adaptive graph fusion. *Pattern Recognition*, pp. 113856, 2026.

Ming Yin, Xin Liu, Junli Gao, Haoliang Yuan, Taisong Jin, Shengwei Zhang, and Lingling Li. Mvaibnet: Multiview disentangled representation learning with information bottleneck. *IEEE Transactions on Industrial Informatics*, 20(10):11511–11520, 2024.

Honglin Yuan, Xingfeng Li, Jian Dai, Xiaojian You, Yuan Sun, and Zhenwen Ren. Deep streaming view clustering. In *Proceedings of the International Conference on Machine Learning*, 2025.

Chao Zhang, Haoxing Chen, Huaxiong Li, and Chunlin Chen. Learning latent disentangled embeddings and graphs for multi-view clustering. *Pattern Recognition*, 156:110839, 2024a.

Chao Zhang, Deng Xu, Xiuyi Jia, Chunlin Chen, and Huaxiong Li. Continual multi-view clustering with consistent anchor guidance. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 5434–5442, 2024b.

Lei Zhang, Lele Fu, Tong Wang, Chuan Chen, and Chuanfu Zhang. Mutual information-driven multi-view clustering. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 3268–3277, 2023.

Bo Zhong, Pengyuan Li, Zisen Kong, Dongxia Chang, Yiming Wang, and Yao Zhao. Anchor-based disentanglement framework for incremental multi-view clustering. *Neural Networks*, pp. 109327, 2026.

Peng Zhou, Yi-Dong Shen, Liang Du, Fan Ye, and Xuejun Li. Incremental multi-view spectral clustering. *Knowledge-Based Systems*, 174:73–86, 2019.

Guoliang Zou, Shizhe Hu, Sijia Li, Yuyang Du, Tongji Chen, and Yangdong Ye. Reliable continual multi-modal clustering. *Pattern Recognition*, 180:114527, 2026.

# A Appendix

This supplementary material provides the complete optimization process, additional experimental details, and extended related work to support the main manuscript. Specifically, Section B presents the complete optimization algorithm of DSMVC along with its complexity analysis. Section C describes the benchmark configurations and reports extended evaluations, including runtime comparisons, ablation studies, and parameter sensitivity. Finally, Section D provides an extended review of related work on disentangled representation learning and streaming multi-view clustering, elaborating on the key differences with DSMVC.

# B Optimization Process and Complexity Analysis

## B.1 Optimization Process of DSMVC

After the sequential knowledge accumulation across all views, DSMVC obtains the unified clustering representation $\mathbf{Z}^* = [\mathbf{Z}_c^1; \mathbf{Z}_s^1; \ldots; \mathbf{Z}_c^V; \mathbf{Z}_s^V] \in \mathbb{R}^{N \times 2Vd}$ by integrating the learned dominant consensus structures and the purified residual variations, where the latter complement shared semantics with view-private diversity. Finally, standard $K$-means clustering is performed on $\mathbf{Z}^*$ to generate the final cluster assignments $\mathbf{y}$. The overall optimization process of DSMVC is summarized in Algorithm 1.

**Algorithm 1:** The optimization process of DSMVC

**Input:** Streaming view sequence $\{\mathbf{X}^v\}_{v=1}^V$; Number of epochs $E$; Trade-off parameters $\lambda_1, \lambda_2$.
**Output:** Final cluster assignments $\mathbf{y}$.

```
Initialize semantic buffer B ← ∅;
for v = 1 to V do
    for e = 1 to E do
        for each mini-batch in X^v do
            Decompose latent features into dominant Z_c^v and residual Z_s^v via Eq. (7);
            Compute reconstruction loss L_r^v via Eq. (9) and suppression loss L_s^v via Eq. (14);
            if v > 1 then
                Compute discrepancy-guided cross-view alignment loss L_c^v using B via
                 Eq. (18);
            else
                L_c^v ← 0;
            Update network parameters using the joint loss L^v in Eq. (19);
    Update historical semantic buffer B ← B ∪ {(Z_c^v, Z_s^v)};
Concatenate multi-view representations Z* = [Z_c^1; Z_s^1; ...; Z_c^V; Z_s^V];
Execute K-means clustering on Z* to obtain final cluster labels y.
```

### B.2 Complexity Analysis

The per-epoch computational complexity of DSMVC is governed by three core modules: Disentangled Feature Learning, Attribution-Guided Suppression Module (AGSM), and Discrepancy-Guided Cross-view Alignment Module (DGAM). First, autoencoder optimization requires $\mathcal{O}(VBM^2)$ for batch size $B$ across $V$ views, where $M = 1024$ denotes the maximum layer width. Additionally, batch-wise SVD on the latent features contributes $\mathcal{O}(VBd^2)$, where $d \ll B$. Second, AGSM evaluates first-order reconstruction sensitivities via a single decoder backward pass, incurring $\mathcal{O}(VBM^2)$ per batch. Third, DGAM aligns the arriving view with $v-1$ historical representations via the predictor network, taking $\mathcal{O}(V^2BdM)$ across all streaming rounds where $\sum_{v=2}^{V}(v-1) = \mathcal{O}(V^2)$. Summing over all $N/B$ batches, the batch size $B$ scales to the full sample size $N$, yielding an overall per-epoch complexity of $\mathcal{O}(VNM^2 + VNd^2 + V^2NdM)$. Since $V, d, M \ll N$, the computational cost scales strictly linearly with sample size $N$, confirming the scalability of DSMVC on large-scale datasets.

## C Additional Experiments

### C.1 Running Time and Efficiency Comparison

To evaluate computational efficiency, we compare the per-epoch training time, inference time, and clustering accuracy of DSMVC against four representative methods on the Handwritten and Out-Scene datasets. As summarized in Table 3, DSMVC achieves the lowest training cost while delivering the highest clustering accuracy across both benchmarks. Specifically, DSMVC runs over $7\times$ faster than the fastest sequence-access algorithm CMSA on Handwritten (389.77 ms vs. 2821.97 ms/epoch), while outperforming the second-best clustering method MGCMVC by 8.48% in ACC. Similarly, Out-Scene evaluations reveal a larger efficiency advantage: our model finishes each epoch in 278.18 ms, surpassing DSVC (2883.82 ms) by more than $10\times$. Although DSMVC exhibits slightly higher latency for inference efficiency, it still operates at a competitive millisecond scale (11.12 ms on Handwritten and 5.13 ms on Out-Scene). These results demonstrate that our DSMVC achieves a superior trade-off between computational efficiency and clustering quality.

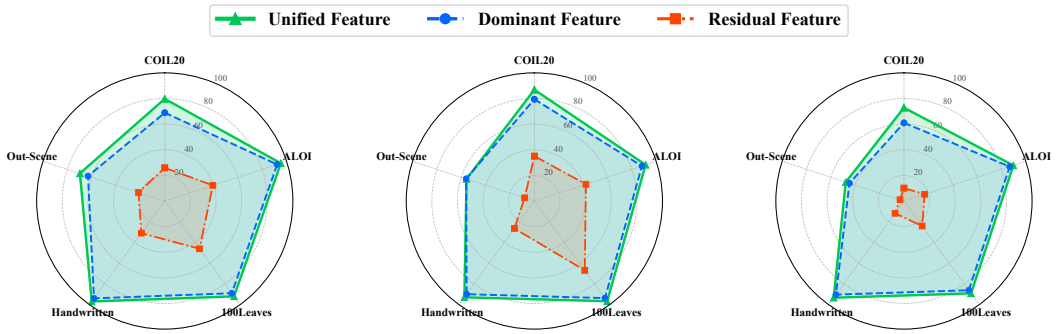
### C.2 Effectiveness of Disentangled Components

To validate the effectiveness of disentangled representation learning, we evaluate the clustering performance of dominant, purified residual, and unified features across five benchmark datasets. As illustrated in Figure 9, the dominant features consistently achieve competitive clustering results across

Table 3: Per-epoch computational efficiency and clustering accuracy across benchmarks. The best and second-best results are highlighted in **bold** and underlined, respectively.

| Method | Handwritten | | | Out-Scene | | |
|---|---|---|---|---|---|---|
| | Training Time (ms) ↓ | Inference Time (ms) ↓ | ACC (%) ↑ | Training Time (ms) ↓ | Inference Time (ms) ↓ | ACC (%) ↑ |
| MGCMVC | 3699.27 | 972.67 | 88.19 | 7233.82 | 2458.29 | 66.37 |
| DFL-Net | 3484.47 | 1408.08 | 64.08 | 6397.00 | 2951.24 | 53.83 |
| DSVC | 3145.81 | **5.48** | 86.90 | 2883.82 | **4.17** | 58.95 |
| CMSA | 2821.97 | 1945.18 | 45.76 | 8855.53 | 287.09 | 39.49 |
| DSMVC | **389.77** | 11.12 | **96.67** | **278.18** | 5.13 | **69.63** |

all metrics, demonstrating that the leading spectral subspace effectively captures the primary semantic structure. In contrast, evaluating residual features alone yields lower yet non-trivial performance, confirming that AGSM filters out detrimental noise while preserving valid view-specific patterns. Most crucially, the unified representation outperforms the dominant branch across all benchmarks, particularly in ACC and ARI. This result confirms that the purified residuals contribute valuable complementary diversity rather than interfering noise, justifying our strategy of integrating both components for final clustering.

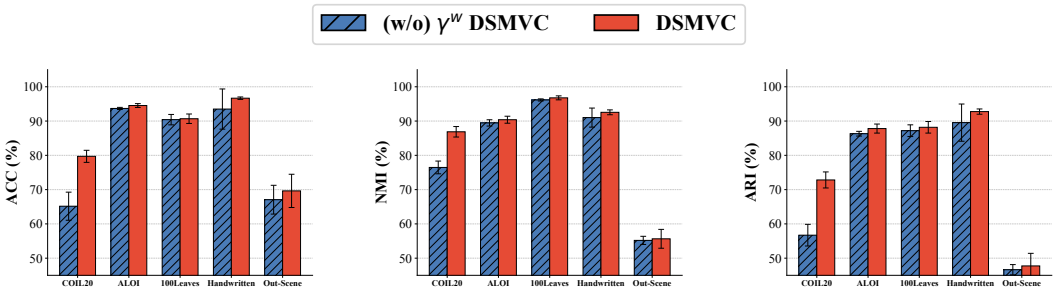



(a) ACC (b) NMI (c) ARI

Figure 9: Performance on five datasets using dominant, residual, and unified features, respectively.

### C.3 Impact of Adaptive Alignment Weights

We conduct an ablation experiment on the adaptive weights $\gamma^w$ in DGAM to verify their impact on evaluating the reliability of cross-view correspondence. Specifically, we compare DSMVC with a variant removing $\gamma^w$, which enforces uniform cross-view associations by setting all pair weights to 1. As shown in Figure 10, DSMVC consistently achieves superior performance and lower variance across all datasets. Notably, the performance degradation without $\gamma^w$ is evident on challenging benchmarks like COIL20 and Out-Scene, suggesting that treating all correspondences equally is disrupted by unreliable associations. In contrast, our method adaptively downweights unreliable pairs via permutation-induced prediction inconsistency, preserving semantic consensus and enhancing representation stability during streaming knowledge accumulation.

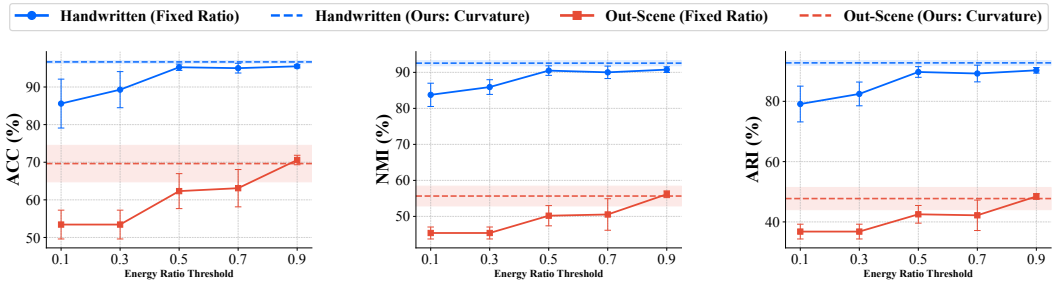



(a) ACC (b) NMI (c) ARI

Figure 10: Ablation study on $\gamma^w$ across five datasets.

### C.4 Evaluation of Curvature-Based Partition Strategy

To assess the curvature-based partition strategy, we benchmark it against fixed energy ratio thresholds on the Handwritten and Out-Scene datasets. As illustrated in Figure 11, setting a rigid energy ratio incurs notable performance discrepancies across different values: small thresholds truncate semantic structure into the residual branch, while excessively large thresholds may leak unstructured noise into the dominant components. Conversely, our adaptive curvature-based partition strategy consistently exceeds or matches the best manual ratios across all evaluated metrics. This indicates that calculating the maximum curvature point reliably identifies the intrinsic elbow where semantic energy approximately saturates, providing a parameter-free and dataset-agnostic truncation boundary between dominant and residual components without manual tuning.

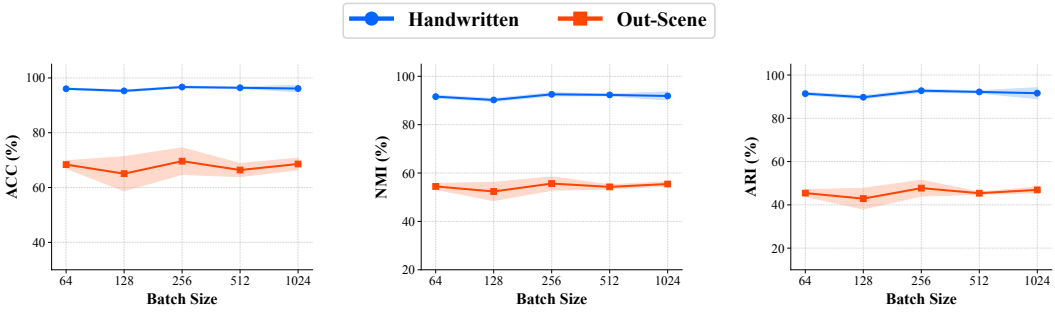



(a) ACC (b) NMI (c) ARI

Figure 11: Sensitivity analysis of the energy ratio threshold on Handwritten and Out-Scene datasets. The dashed lines represent our parameter-free curvature truncation method.

### C.5 Stability Analysis across Mini-batch Sizes

We investigate the sensitivity of DSMVC across different mini-batch sizes to assess the stability of SVD decomposition during mini-batch training. As illustrated in Figure 12, DSMVC maintains consistently stable clustering performance across a wide range of batch sizes on both datasets. Even under relatively small batch sizes (e.g., Batch Size = 32), the proposed method still exhibits minimal performance fluctuation. This empirical stability confirms that the spectral energy distribution within mini-batches can reliably capture the dominant low-rank geometric structures of incoming views, thereby avoiding severe boundary oscillations between the dominant and residual subspaces. Consequently, our curvature-based partition strategy proves robust to mini-batch training, ensuring practical scalability and reliable disentanglement without relying on batch sizes.



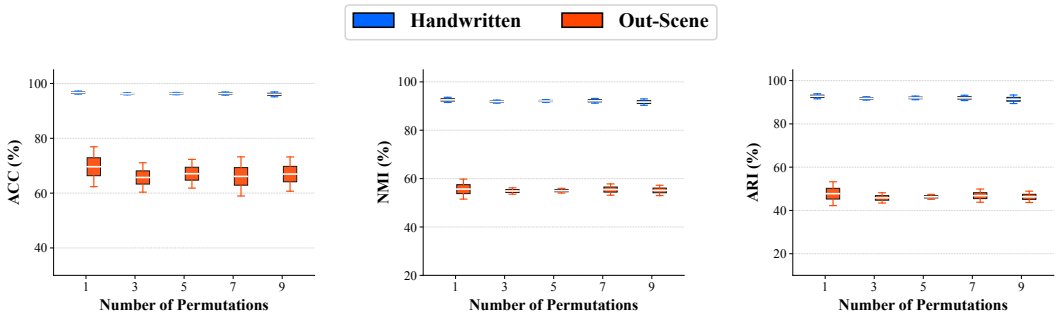

(a) ACC (b) NMI (c) ARI

Figure 12: Clustering performance across different mini-batch sizes on Handwritten and Out-Scene datasets. The shaded regions denote the standard deviations across multiple runs.

### C.6 Robustness to Permutation Counts in DGAM

Figure 13 demonstrates the clustering performance across different permutation numbers in DGAM on both benchmarks. Specifically, these results exhibit steady clustering quality with negligible

variance as the permutation count increases, indicating that the proposed method is insensitive to the number of permutation counts. While averaging multiple random permutations narrows the dispersion band on challenging distributions, a single permutation sufficiently disrupts authentic cross-view correspondences to establish an effective error baseline. Consequently, relying on a single shuffled baseline provides a reliable calibration signal for cross-view alignment, preserving both computational efficiency and high clustering quality during streaming accumulation.




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

Figure 13: Clustering performance across different numbers of DGAM permutations on Handwritten and Out-Scene datasets. The white lines inside the boxes indicate the mean performances.

### C.7 Benchmark Datasets and Implementation Setup

We evaluate DSMVC across seven public multi-view benchmarks spanning diverse modalities and scales: COIL20 (Nene et al., 1996), ALOI (Geusebroek et al., 2005), 100Leaves (Mallah et al., 2013), Handwritten (Breukelen et al., 1998), Out-Scene (Oliva & Torralba, 2001), ALOI100 (Geusebroek et al., 2005), and Cifar100 (Wu et al., 2024). The comprehensive data statistics, including sample sizes, cluster counts, view numbers, and dimensionalities, are summarized in Table 4.

To extract view-specific latent representations, we equip a three-layer encoder with dimensions $[D^v, 1024, 1024, d]$ for each view, while the corresponding decoder follows a symmetric structure of $[2d, 1024, 1024, D^v]$ for reconstruction, where $D^v$ and $d$ denote the dimensions of the raw input features and the unified embedding space, respectively. Moreover, an additional prediction network with dimensions $[d, 1024, d]$ is utilized for cross-view knowledge transfer. Furthermore, the entire framework is optimized using the Adam optimizer with a learning rate of $0.0001$, and the embedding dimension, batch size, and training epochs are set to 256, 256, and 100, respectively. All experiments are implemented in Python 3.9.19 and PyTorch 1.12.1 with CUDA 11.3 on an Ubuntu 18.04.5 LTS server equipped with an NVIDIA TITAN Xp GPU (12GB RAM).

Table 4: Detailed statistics of the benchmark multi-view datasets.

| Dataset | Samples ($N$) | Clusters ($K$) | Views ($V$) | View Dimensions ($D^v$) |
|---|---|---|---|---|
| COIL20 | 480 | 20 | 3 | 4096 / 4096 / 4096 |
| ALOI | 1079 | 10 | 4 | 64 / 64 / 77 / 13 |
| 100Leaves | 1600 | 100 | 3 | 64 / 64 / 64 |
| Handwritten | 2000 | 10 | 6 | 216 / 76 / 64 / 6 / 240 / 47 |
| Out-Scene | 2688 | 8 | 4 | 512 / 432 / 256 / 48 |
| ALOI100 | 10800 | 100 | 4 | 77 / 13 / 64 / 125 |
| Cifar100 | 50000 | 100 | 3 | 512 / 2048 / 1024 |

## D Related Work

### D.1 Disentangled Multi-view Clustering

Disentangled multi-view clustering aims to decompose shared consensus from view-specific variations to eliminate redundancy and enhance clustering discriminability (Ke et al., 2024; 2023;

Wang et al., 2024; Yan et al., 2024; Zhang et al., 2023). Early efforts mainly adopt generative frameworks such as Multi-VAE (Xu et al., 2021) and MVAIBNet (Yin et al., 2024), which formulate common and private factors via variational objectives or the information bottleneck principle. However, these methods often struggle with posterior collapse, prompting discriminative approaches to enforce explicit statistical independence. For example, DFL-Net (Chen et al., 2025) and DMVCS (Li et al., 2026) leverage contrastive objectives to isolate private factors, while DC-SPAN (Chen et al., 2026) introduces dual-path architectures for shared-private subspace disentanglement. Furthermore, to capture high-order latent dependencies beyond pairwise relations, structured methods like DSTL (Xu et al., 2024) and DEGREE (Zhang et al., 2024a) introduce tensor-based low-rank regularization on disentangled representations or view-specific graphs. Despite this progression, existing frameworks predominantly treat view-specific spaces as homogeneous entities through holistic penalties or coarse-grained component separation. In practice, view-specific representations interweave beneficial complementarity with detrimental variations. Indiscriminately suppressing or aligning these spaces corrupts consensus learning through noise interference. Consequently, DSMVC employs a curvature-driven partition strategy with attribution-based sensitivity suppression, purifying residual variations to preserve fine-grained complementary semantics.

### D.2 Streaming Multi-view Clustering

Streaming Multi-view Clustering (SMVC) coordinates cross-view knowledge accumulation across sequentially arriving modalities under strict memory constraints, preventing catastrophic forgetting while updating the unified semantic space (Li et al., 2024; Sun et al., 2020). Recent studies have advanced along two main directions: Prototype-based Alignment and Dynamic Replay Learning (Feng et al., 2025; Zou et al., 2026). Prototype-based alignment methods focus on maintaining compact historical summaries to guide the representation learning of newly arrived views. A representative method is DSVC (Yuan et al., 2025), which constructs historical semantic prototypes to alleviate concept drift and align streaming modalities. Similarly, CMVC-PCAL (Wang & Lu, 2026) integrated pseudo-labels with anchor-guided temporal consistency to calibrate incoming features against cached cluster centroids. Conversely, Dynamic replay learning methods aim for continuous parameter consolidation and structural preservation across sequential updates without storing raw data. Following this intuition, CKR-MVC (Liang et al., 2025) coupled contrastive replay with structural distillation to sustain historical manifold geometries over streaming rounds. Furthermore, to constrain cross-stream interactions, ADIMC (Zhong et al., 2026) restricted knowledge transfer to shared subspaces. While these methods have achieved promising performance, they typically enforce cross-view alignment across views, implicitly assuming that all pairwise associations between historical and current views are equally reliable. In contrast, DSMVC introduces discrepancy-guided cross-view alignment to adaptively emphasize reliable cross-view correspondences, reducing the influence of weakly supported associations and thus promoting stable consensus.