SparseMVC: Probing Cross-view Sparsity Variations for Multi-view Clustering

Ruimeng Liu[1], **Xin Zou**[2], **Chang Tang**[3*]
Xiao Zheng[4], **Xingchen Hu**[5], **Kun Sun**[1], **Xinwang Liu**[5]
[1]School of Computer Science, China University of Geosciences (Wuhan)
[2]The Hong Kong University of Science and Technology (Guangzhou)
[3]School of Software Engineering, Huazhong University of Science and Technology
[4]School of Computer Science, Hubei University of Technology
[5]College of Systems Engineering, National University of Defense Technology

Abstract

Existing multi-view clustering methods employ various strategies to address data-level sparsity and view-level dynamic fusion. However, we identify a critical yet overlooked issue: *varying sparsity across views*. Cross-view sparsity variations lead to encoding discrepancies, heightening sample-level semantic heterogeneity and making view-level dynamic weighting inappropriate. To tackle these challenges, we propose Adaptive **Sparse** Autoencoders for **M**ulti-**V**iew **C**lustering (SparseMVC), a framework with three key modules. Initially, the sparse autoencoder probes the sparsity of each view and adaptively adjusts encoding formats via an entropy-matching loss term, mitigating cross-view inconsistencies. Subsequently, the correlation-informed sample reweighting module employs attention mechanisms to assign weights by capturing correlations between early-fused global and view-specific features, reducing encoding discrepancies and balancing contributions. Furthermore, the cross-view distribution alignment module aligns feature distributions during the late fusion stage, accommodating datasets with an arbitrary number of views. Extensive experiments demonstrate that SparseMVC achieves state-of-the-art clustering performance. Our framework advances the field by extending sparsity handling from the data-level to view-level and mitigating the adverse effects of encoding discrepancies through sample-level dynamic weighting. The source code is publicly available at https://github.com/cleste-pome/SparseMVC.

1 Introduction

Multi-view learning has emerged as a powerful paradigm for leveraging complementary information across multiple perspectives, significantly improving the performance of unsupervised learning tasks such as clustering [1, 2, 3, 4, 5]. At the same time, the sparsity of multi-view data has become a pivotal focus of research, with numerous studies proposing solutions from perspectives such as activation functions [6], tensor decomposition [7, 8], and variational autoencoders [9]. Nevertheless, while prior methods focus on designing advanced approaches to address data sparsity, they often overlook a fundamental aspect—the potential variations in sparsity across different views. Given that multi-view data consists of multiple views originating from distinct sources, and sparsity is a pervasive characteristic in multi-view data [10, 11, 12, 13, 14]. Building upon the factual observations, a natural question arises: "*Does there exist a phenomenon of varying sparsity across views?*"

*Corresponding author: Chang Tang (tangchang@hust.edu.cn).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

To quantify cross-view sparsity variations, we define the sparsity ratio $\boldsymbol{s}_v$ for the v-th view:

$$s_v = \frac{1}{N \cdot F} \sum_{j=1}^{N} \sum_{i=1}^{F} \boldsymbol{I}[x_{i,j}^v = 0], \tag{1}$$

where N refers to the number of samples, F refers to the feature dimension, $x_{i,j}^v$ represents the i-th feature of the j-th sample in the v-th view, and the indicator function $\boldsymbol{I}$ takes the value of one if $x_{i,j}^v$ equals zero, and zero otherwise. Zero-valued features $x_{i,j}^v$ suggests missing dimensions or data collection errors. Our statistical and computational analysis addresses the question posed earlier and reveals that sparsity variations across views not only exist, but are widely prevalent in diverse multi-view data, as illustrated in Fig. 1.

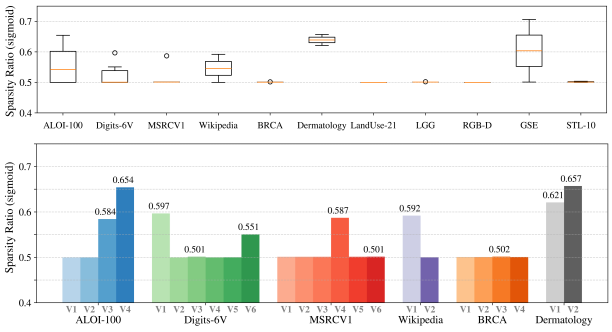


Figure 1: Sparsity ratios across views in multi-view datasets. TOP BOX PLOT illustrates the sparsity ratio distribution, which shows the median (orange line), interquartile range (box), and any outliers (points outside the whiskers). BOTTOM BAR PLOT presents the sparsity ratios for each view within each dataset. To provide a more comprehensive situation of sparsity variations, additional datasets are included: Digits [15], LandUse-21 [16], RGB-D [17]. GSE [18] and STL-10 [19]. The cross-view sparsity ratios have been processed to improve visualization using **the sigmoid function**.

The disparity in view sparsity presents a multifaceted challenge: highly sparse views, lacking sufficient informative content, are prone to underfitting, whereas less sparse views, often burdened with redundant or irrelevant features, are susceptible to overfitting. Applying a uniform encoder architecture or regularization strategy across such heterogeneous views compromises representational consistency and limits the model's capacity to extract complementary cross-view information. To resolve this, we adopt an adaptive and sparsity-aware encoding strategy tailored to individual views. This requires rethinking the autoencoder design to accommodate view heterogeneity and structural disparities. Our solution is to design an autoencoder capable of adaptively adjusting its constraints based on the sparsity ratio of each view, allowing its encoding form to evolve accordingly.

Due to varying sparsity across views for the same sample, encoders can introduce semantic and representational inconsistencies, affecting subsequent stages [20, 21]. As a result, it becomes essential to dynamically assess the contribution of each view based on the features extracted from different types of encoders. To address this, we design a correlation-informed reweighting module that assigns sample-wise weights based on the correlations between global and local features, thereby balancing view contributions and discrepancies. The early fusion strategy integrates multi-view data through feature concatenation. Since our reweighting module uses the global latent representation encoded from early fusion to guide the subsequent dynamic weighted early fusion of local representations, we chose an early fusion approach that preserves the original feature values as much as possible.

In late fusion, we introduce a cross-view distribution alignment module to align feature distributions across views, enabling robust integration and supporting datasets with arbitrary view numbers while balancing global consistency and view-specific diversity. These three core components together form the SparseMVC framework, addressing sparsity inconsistencies, encoding discrepancies, and semantic heterogeneity in a step-by-step and interdependent design.

Table 1: Performance comparison of different strategies under extreme and minimal sparsity variation.

Datasets	**Sparsity of Different Views**	**Accuracy** [Ours]	[2nd]	**NMI** [Ours]	[2nd]
ALOI-100	[0.0001, 0.0001, 0.3415, 0.6383]	82.21 ↑4.49	77.72	92.65 ↑1.78	90.87
MSRCV1	[0.0049, 0.0048, 0.0048, 0.3478, 0.0051, 0.0048]	97.14 ↑5.71	91.43	94.22 ↑6.32	87.90
LGG	[0.0040, 0.0038, 0.0078, 0.0037]	83.15 ↑0.38	82.77	54.62 ↑0.49	54.13
Synthetic3d	[0.0017, 0.0017, 0.0017]	98.33 ↑0.16	98.17	92.01 ↑0.74	91.27

As presented in Table 1, we evaluate our method across a range of datasets exhibiting extreme sparsity disparities. For instance, in ALOI-100, the sparsity ratio spans from as low as 0.0001 to as high as 0.6644, revealing a substantial imbalance in information density across views. To effectively

address such heterogeneity, our framework incorporates dynamically adaptive autoencoders that adjust both the encoding process and sparsity-aware regularization in accordance with each view's sparsity profile. Unlike existing methods, which generally overlook the impact of cross-view sparsity variation, our targeted design enables SparseMVC to achieve consistently superior performance under severe disparity conditions. Moreover, it maintains strong competitiveness even in datasets with near-uniform sparsity variations, such as LGG and Synthetic3D, demonstrating both its responsiveness to sparsity imbalance and its robustness in more homogeneous scenarios.

To the best of our knowledge, this is the first work to explicitly identify, analyze, and define the problem of cross-view sparsity variations in multi-view data, and to propose a dedicated framework SparseMVC that offers a targeted and principled solution.

2 Related Work

For multi-view fusion, early methods [22, 23] assumed equal importance of views, ignoring view heterogeneity. Hence, dynamic weighting approaches have emerged, with attention-based methods [24, 25] leading, alongside loss optimization [26, 27], kernel techniques [28, 29], and subspace methods [30]. However, uniform view-level weighting fails to address intra-view variability, highlighting the need for sample-level dynamic weighting. Trust-based methods [31, 32, 33, 34] excel in supervised scenarios. For multi-view clustering, sample-adaptive fusion [35] has been proposed using Laplacian matrix divergence. In contrast, our framework dynamically computes sample weights via correlation calculation without additional loss, while adopting a decoupled design with independent global and view-specific autoencoders. More details can be found in Appendix A.2.

Sparse representation effectively captures essential features in sparse data by enforcing sparsity constraints [36] but struggles to handle the non-linear structures common in multi-view datasets [37]. Autoencoders excel at learning non-linear latent features [38, 39], yet their lack of sparsity enforcement limits their adaptability to varying sparsity rates across views. Sparse autoencoders combine the strengths of both sparse representation and standard autoencoders by incorporating sparsity constraints into the hidden layers [40], which have become a research focus in multi-view learning [9, 41].

3 Method

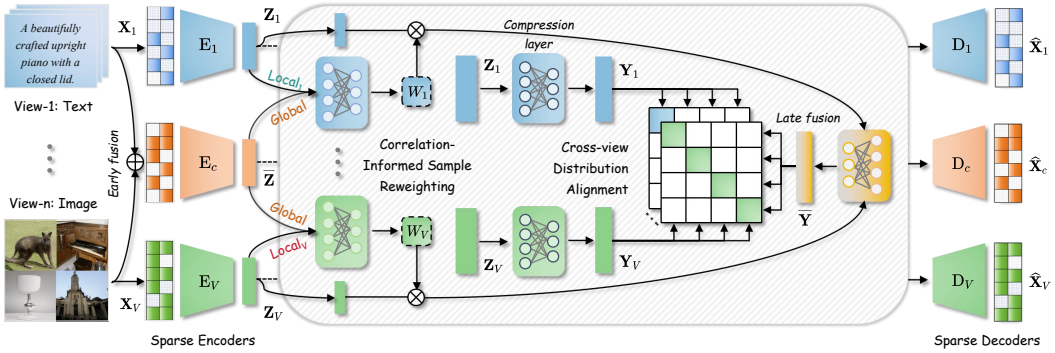


Figure 2: Overview of **SparseMVC**, a framework designed to address varying sparsity across views.

This section sequentially introduces the three key submodules of SparseMVC, as illustrated in Fig. 2. It begins by utilizing sparse autoencoders with adaptive constraints, which dynamically adjust the coding strategy based on the probed s_v, to generate latent features ($\boldsymbol{Z}$), making the reconstructed features ($\widehat{\boldsymbol{X}}$) approximate the original input features ($\boldsymbol{X}$). Subsequently, the correlation between the early-fused global features ($\bar{\boldsymbol{Z}}$) and view-specific features ($\{\boldsymbol{Z}_v\}_{v=1}^n$) guides the computation of sample-level weights ($\{\boldsymbol{W}_v\}_{v=1}^n$) via the attention mechanism within the correlation-informed sample reweighting module. Finally, the cross-view distribution alignment module enhances clustering performance by setting the late-fused global features $\overline{\boldsymbol{Y}}$ as the anchor latent representation, and then simultaneously aligning the multi-view feature distribution between $\overline{\boldsymbol{Y}}$ and each view-specific compressed feature ($\{\boldsymbol{Y}_v\}_{v=1}^n$). The algorithm of the framework can be found in Appendix A.1.

3.1 Sparse Autoencoder with Adaptive Constraints

To handle varying view sparsity rates, we propose the sparse autoencoder with adaptive constraints (SAA), extending traditional sparse autoencoders. SAA employs an adaptive loss function that integrates reconstruction and sparsity-aware entropy-matching as distinct constraints, wherein the adjustment is dynamically guided by view sparsity ratios formulated as prior knowledge.

The reconstruction loss, typically measured by mean squared error (MSE), quantifies the difference between the reconstructed output $\hat{\boldsymbol{x}}_j^v$ and the input $\boldsymbol{x}_j^v$ for the j-th sample:

$$\mathcal{L}_{\text{recon}}^v = \frac{1}{N} \sum \left(\hat{\boldsymbol{x}}_j^v - \boldsymbol{x}_j^v \right)^2, \tag{2}$$

where N is the number of samples (batch size) for view v. Motivated by the widespread presence of sparsity variations in different views, our aim is to design a function that is positively correlated with s_v, allowing adaptive adjustments to both the encoder type and the strength of the sparsity constraints. We scale by $f(s_v)$ outside the loss rather than tuning ρ in Eq. (4), please refer to Appendix B.2. The design of the adaptive weighting factor $f(s_v)$ follows a ReLU-like approach, adjusting the strength of $\mathcal{L}_{\text{entropy}}^v$ based on the probed input dataset sparsity s_v for view v:

$$f(\boldsymbol{s}_v) = \begin{cases} 0, & \text{if } \boldsymbol{s}_v \leq \theta, \\ \frac{\boldsymbol{s}_v - \theta}{1 - \theta}, & \text{if } \boldsymbol{s}_v > \theta, \end{cases} \tag{3}$$

where the default value of θ is 0.01. Selecting the threshold θ for the sparsity ratio s_v is based on the actual cross-view sparsity distribution of each dataset. For most datasets, the sparsity ratios exhibit a skewed distribution, with a significant concentration of both high and low values, as shown in Fig. 1 and Table 1. A θ value of 0.01 effectively captures these low-sparsity views. To enforce sparsity constraints, the entropy-matching loss is defined using Kullback-Leibler (KL) [42] divergence, encouraging the activations $\hat{\boldsymbol{h}}_{v,k}$ in the hidden layer to align with a target sparsity level ρ. The entropy-matching loss and the sparsity loss derived from it are formulated as follows:

$$\mathcal{L}_{\text{sparse}}^v = f(s_v) \cdot \mathcal{L}_{\text{entropy}}^v = f(s_v) \cdot \sum_{k=1}^{H} (\rho \log \frac{\rho}{\hat{\boldsymbol{h}}_k^v} + (1-\rho) \log \frac{1-\rho}{1-\hat{\boldsymbol{h}}_k^v}), \tag{4}$$

where H refers to the number of units in the first hidden layer of the sparse encoder, ρ is the target sparsity level, and $\hat{\boldsymbol{h}}_k^v$ represents the average activation of the k-th hidden unit for view v. Following [40], ρ is set to 0.05, which is a well-validated choice that balances sparsity and the learning capacity of the autoencoder, allowing it to effectively capture key features in the data while avoiding overfitting to irrelevant features. The average activation is computed by

$$\hat{\boldsymbol{h}}_k^v = \frac{1}{N} \sum_{j=1}^{N} \sigma \left(\boldsymbol{W}_k^v \boldsymbol{x}_j^v + \boldsymbol{b}_k^v \right), \tag{5}$$

where $\boldsymbol{W}_k^v$ is the weight matrix, $\boldsymbol{b}_k^v$ is the bias term for the k-th hidden unit in the v-th view, $\boldsymbol{x}_j^v$ is the input feature, and $\sigma(\cdot)$ denotes the ReLU activation function. When $s_v \leq \theta$, $\mathcal{L}_{\text{entropy}}^v$ is deactivated ($f(\boldsymbol{s}_v) = 0$) and the sparse autoencoder degenerates into a standard autoencoder. Conversely, for input views where $\boldsymbol{s}_v > \theta$, $f(\boldsymbol{s}_v)$ exhibits a linear increase with $\boldsymbol{s}_v$, ensuring that the sparsity constraint becomes more prominent for highly sparse inputs. Based on the above, the loss function for SAA is then formulated as follows:

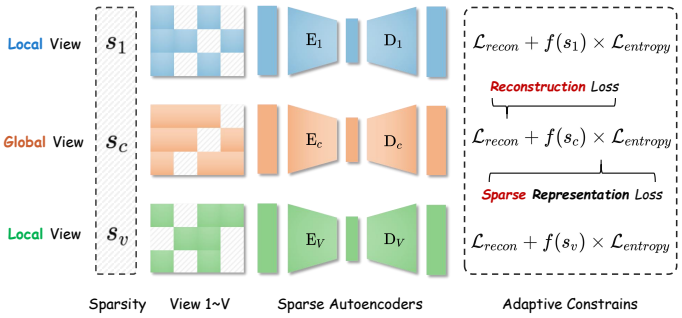


Figure 3: Sparse autoencoder with adaptive constraints.

$$\mathcal{L}_{\text{SAA}} = \sum_{v=1}^{V} \left(\mathcal{L}_{recon}^v + \mathcal{L}_{\text{sparse}}^v \right). \tag{6}$$

As shown in Fig. 3, this piecewise linear design dynamically aligns the sparsity constraint with the input sparsity level and customizes the encoding strategy for each view.

3.2 Correlation-Informed Sample Reweighting

While SAA balances reconstruction and sparsity, it introduces sample-specific encoding inconsistencies due to differences in sparsity. Additionally, while networks adjust weights through layer updates, this weighting alone cannot address the lack of communication between autoencoders. On the above bases, we drew inspiration from the concept of multi-head attention and designed the correlation-informed sample reweighting module (CSR), which leverages correlations between the early-fused global features ($\bar{\boldsymbol{Z}}$) and the view-specific local features ($\boldsymbol{Z}_v$), then computes sample-specific weights. This cascading design is devised to achieve two objectives: mitigating the encoding inconsistencies introduced by SAA and leveraging globally fused features, which preserve the relatively high-fidelity patterns of the original data structure, to supervise the computation of view correlations.

CSR adopts a simplified structure inspired by multi-head attention, which captures similar effects while deviating from the standard formulation. Initially, CSR takes $\bar{\boldsymbol{Z}} \in \mathbb{R}^{N\times F}$ and $\boldsymbol{Z}_v \in \mathbb{R}^{N\times F}$ as input, and projects them into the query, key, and value spaces through parallel linear transformations:

$$\boldsymbol{Q} = \bar{\boldsymbol{Z}}\boldsymbol{W}_Q,\ \boldsymbol{K}_v = \boldsymbol{Z}_v\boldsymbol{W}_K,\ \boldsymbol{V}_v = \boldsymbol{Z}_v\boldsymbol{W}_V, \tag{7}$$

where $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{F\times F}$ are the learnable weight matrices, generating the query matrix $\boldsymbol{Q} \in \mathbb{R}^{N\times F}$, which encapsulates global semantic information, and the key matrix $\boldsymbol{K}_v \in \mathbb{R}^{N\times F}$, which capture view-specific features for each view. Our primary objective is to quantify inter-view relationships by evaluating attention scores between queries and keys, without generating new feature representations, thus omitting the value matrix $\boldsymbol{V}_v$. To compute the correlation between $\bar{\boldsymbol{Z}}$ and $\boldsymbol{Z}_v$, we define the correlation score $\boldsymbol{C}_v \in \mathbb{R}^N$ for the v-th view based on Einstein summation convention as:

$$\boldsymbol{C}_v = \frac{\sum_{f=1}^{F} \boldsymbol{Q}_f \cdot (\boldsymbol{K}_f^v)^T}{\sqrt{F}}, \tag{8}$$

where $\sqrt{F}$ is the scaling factor equals to the square root of the dimension of the key vector. The correlation scores are normalized via the softmax function to produce sample-specific $\boldsymbol{W}_v \in \mathbb{R}^N$ as:

$$\boldsymbol{W}_v = \frac{\exp(\boldsymbol{C}_v)}{\sum_{v=1}^{V} \exp(\boldsymbol{C}_v)}, \tag{9}$$

that dynamically adjust the contribution of each corresponding sample in respective autoencoders.

3.3 Cross-view Distribution Alignment

Aligning features across views is a fundamental challenge in multi-view learning, as it is crucial for leveraging the complementary information provided by diverse views. The cross-view distribution Alignment module (CDA) addresses this issue by performing contrastive learning between the late-fused global features ($\overline{\boldsymbol{Y}}$) and the compressed features of individual views ($\boldsymbol{Y}_v$), ensuring effective alignment of multi-view features within a unified and shared latent space.

To mitigate the risk of dimensional collapse during alignment, potentially caused by an excessively large latent space, we introduce a compression layer before feeding the encoded view-specific features into the CDA. More details are provided in Appendix B.1. Specifically, $\boldsymbol{Y}_v \in \mathbb{R}^{N\times F}$ is obtained from $\boldsymbol{Z}_v$ via the compression layer. In parallel, the global features $\overline{\boldsymbol{Y}} \in \mathbb{R}^{N\times F}$ are as follows:

$$\overline{\boldsymbol{Y}} = \mathcal{F}\left(\sum_{v=1}^{V} \boldsymbol{W}_v \boldsymbol{Z}_v\right), \tag{10}$$

where fusion function $\mathcal{F}$ represents the late fusion layers and ensures that the transformed dimension matches $\boldsymbol{Y}_v$. The similarity matrix $\boldsymbol{S}_v$ between $\overline{\boldsymbol{Y}}$ and $\boldsymbol{Y}_v$ is defined as:

$$\boldsymbol{S}_v = \frac{\overline{\boldsymbol{Y}} \cdot (\boldsymbol{Y}_v)^T}{\tau}, \tag{11}$$

with τ denoting a temperature parameter that scales the similarity values. The sample pairs position indices in $\boldsymbol{S}_v$ are defined as p and q. Positive pairs, which correspond to the same samples across

views , are represented by the diagonal elements of $\boldsymbol{S}_v$, denoted as $\boldsymbol{S}_v^{p,p}$. Negative pairs, which involve different samples across views, are identified using a mask matrix $\boldsymbol{M}_v \in \mathbb{R}^{N\times N}$, where $\boldsymbol{M}_v^{p,q} = 1$ if $p \neq q$, and $\boldsymbol{M}_v^{p,q} = 0$ otherwise. The contrastive loss for each sample is:

$$\mathcal{L}_{con}^{p,v} = -\log\left(\frac{\exp(\boldsymbol{S}_v^{p,p})}{\sum_{q=1}^{N}\exp(\boldsymbol{S}_v^{p,q})\cdot\boldsymbol{M}_v^{p,q}}\right), \tag{12}$$

where $\exp(S^{p,p})$ quantifies the similarity of positive pairs, and the denominator aggregates the exponential similarities of all pairs, weighted by the mask matrix $\boldsymbol{M}_v$. The overall CDA loss across all views is obtained by summing the individual losses for each view and averaging over all samples:

$$\mathcal{L}_{\text{CDA}} = \sum_{v=1}^{V}\frac{1}{N}\sum_{p=1}^{N}\mathcal{L}_{con}^{p,v}. \tag{13}$$

For contrastive learning, when samples of the same class are clustered in one view, the attraction exerted by positive pairs propagates to other views. In contrast, although the distinction between samples does not necessarily imply class disparity, repulsion among negative pairs enables view with greater discriminative power to transmit class separations to other views. This results in a mechanism that transforms both the alignment and misalignment information at the cross-view sample level into an objective, aiming to minimize intra-class while maximizing inter-class distances.

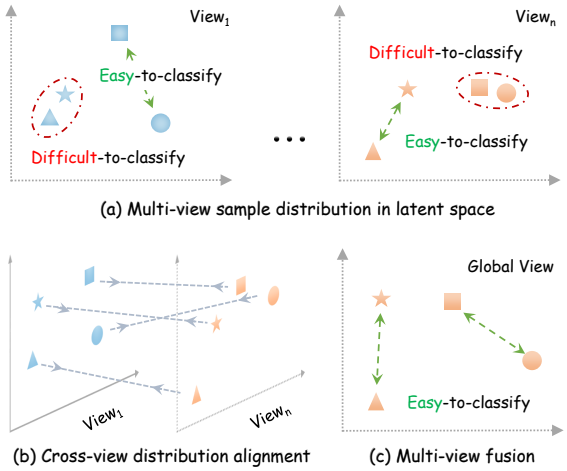
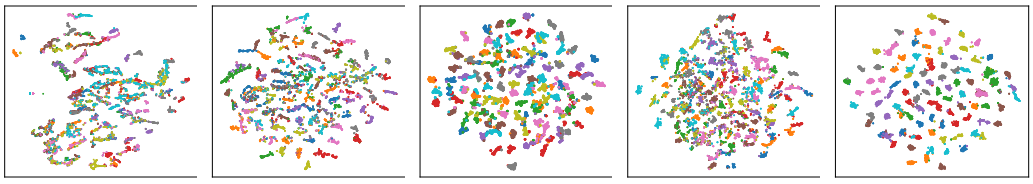


Figure 4: View distribution alignment based on contrast.

Regarding the role of CDA, the global view serves as an anchor, which is compared in parallel against each local view. To minimize the overall contrastive loss, the sample distribution of the global view is concurrently attracted toward all local views, thereby encouraging the features of each sample to converge more tightly in the latent space. From the perspective of the entire latent space, this process effectively facilitates overall distribution alignment. The rationale behind utilizing local views for distribution alignment with the global view to enhance class separability lies in the ability to leverage easily classified samples from one view to improve the distinguishability of harder-to-classify samples in another. The reasoning above, together with Fig. 4, illustrate the working principle of the CDA: utilizing contrastive learning as a tool, through the process of aligning the distribution of sample features across views, enhancing the differentiation of difficult-to-classify samples in one view by leveraging easy-to-classify samples in another, and ultimately achieving the goal of optimal alignment of features in the shared latent space.



(a) GCFAgg (CVPR'23) (b) CPSPAN(CVPR'23) (c) SCMVC (TMM'24) (d) MVCAN (CVPR'24) (e) SparseMVC (Ours)

Figure 5: T-SNE visualization of the features learned with recently comparative methods (a-d) and ours (e) on the ALOI-100 dataset.

3.4 The Overall Loss Function of SparseMVC

The total loss $\mathcal{L}_{\text{total}}$ comprises the adaptive sparse autoencoder loss $\mathcal{L}_{\text{SAA}}$ in Eq. (6), preserving data fidelity and enforcing structured sparsity via $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{entropy}}$, and the cross-view alignment loss

$\mathcal{L}_{\text{CDA}}$ in Eq. (13), ensuring consistent clustering across views:

$$\mathcal{L}_{\text{total}} = \sum_{v=1}^{V} \left(\mathcal{L}_{\text{recon}}^{v} + f(s_v) \cdot \mathcal{L}_{\text{entropy}}^{v} \right) + \lambda_{\text{CR}} \cdot \mathcal{L}_{\text{CDA}}, \tag{14}$$

where λ_{CR} is the constraint ratio coefficient that controls the trade-off between $\mathcal{L}_{\text{SAA}}$ and $\mathcal{L}_{\text{CDA}}$.

4 Experiments

4.1 Experimental Settings

Compared Methods. Our proposed method is compared against the following 12 state-of-the-art multi-view clustering methods based on deep learning. DSMVC [43], COMPLETER [44], DCP [45], CVCL [46] and SCMVC [47] focus on dynamic contrastive learning; MFLVC [48], GCFAgg [3], DealMVC [49] combine feature fusion and consistency; DSMVC [43] and SDMVC [50] enhance consistency through discriminative learning; CPSPAN [51] and MVCAN [52] apply proxy supervision and prototype alignment.

Benchmark Datasets. The selected datasets span diverse domains: Image datasets include MSRCV1 [53] focusing on objects and scenes, Dermatology [54] on medical images, Out-Scene [55] on natural scenes, and ALOI-100 [56] on object recognition. Image-text datasets include Wikipedia [2], which provides website cross-modal data. Omics datasets include LGG [57] focusing on brain tumor genomics and BRCA [58] on breast cancer genomics. Synthetic3d [59] supports 3D object modeling and recognition. Detailed properties of datasets are listed in Table 2.

Table 2: Characteristics of kinds of multi-view datasets.

Datasets	Samples	Clusters	Views	View Dimensions
Images				
MSRCV1	210	7	6	[1302, 48, 512, 100, 256, 210]
Dermatology	358	6	2	[12, 22]
Out-Scene	2,688	8	4	[512, 432, 256, 48]
ALOI-100	10,800	100	4	[77, 13, 64, 125]
Image-Text				
Wikipedia	693	10	2	[128, 10]
Omics				
LGG	267	3	4	[2000, 2000, 333, 209]
BRCA	398	4	4	[2000, 2000, 278, 212]
Synthetics				
Synthetic3d	600	3	3	[3, 3, 3]

Evaluation Metrics. Accuracy (ACC) evaluates alignment with ground truth, normalized mutual information (NMI) measures shared information, purity (PUR) assesses cluster homogeneity. Adjusted Rand index (ARI), measuring clustering similarity, is partially utilized in experiments. For all metrics, higher values indicate better performance.

Implementation Details All experiments were conducted using Python 3.8.15 and PyTorch 1.13.1+cu116 on a Windows PC equipped with an AMD Ryzen 9 5900HX CPU, 32GB RAM, and an Nvidia RTX 3080 GPU (16GB). Models were trained using the Adam optimizer [60], a learning rate of 0.003, and a fixed seed of 50, with batch size equal to the dataset's sample count. Pre-training was performed uniformly for 300 epochs, while alignment training was conducted for 300 epochs for datasets with less than 2500 samples and 1000 epochs for larger datasets. For clustering, k-means [61] was applied with the number of clusters equal to the dataset categories and 100 initializations. During pre-training, global features $\boldsymbol{Z}_v$ derived from early fusion were used, while alignment training used late fusion features $\overline{\boldsymbol{Y}}$. Metrics were calculated as the average of 10 runs in the final epoch, with no fine-tuning performed for specific datasets. To ensure fairness, the hyperparameters for the comparison methods were determined based on either the default global settings or the configuration of the first dataset.

4.2 Comparative Results Analysis

Table 3 and 4 summarize the comparative results, leading to the following conclusions:

(1) Our method achieves state-of-the-art performance across eight diverse multi-view datasets, along with larger-scale datasets as shown in Table 8. These results validate the versatility of our approach and highlight its potential for a wide range of downstream tasks. In comparison, other approaches such as SCMVC, MVCAN, and CPSPAN achieved relatively good results on specific datasets but failed to maintain an advantage due to their limited generalizability across other datasets.

[2]https://dumps.wikimedia.org/

Table 3: Clustering results on small multi-view datasets. The top-ranked result is **bolded**, and the second-ranked result is underlined.

Methods \ Datasets	Synthetic3d			LGG			Dermatology			BRCA		
	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
COMPLETER [CVPR'21] [44]	93.33	76.06	93.33	80.15	49.25	80.15	77.65	80.11	82.12	55.53	34.65	65.33
DCP [TPAMI'22] [45]	97.17	87.60	97.17	59.55	44.82	73.03	72.91	77.22	80.73	57.29	39.51	60.55
MFLVC [CVPR'22] [48]	90.67	72.59	90.67	79.03	49.73	79.03	58.10	56.20	62.85	55.53	27.74	60.05
DSMVC [CVPR'22] [43]	96.83	86.64	96.83	82.77	54.13	82.77	92.74	87.82	92.74	54.52	33.53	68.84
SURE [TPAMI'22] [62]	96.33	85.16	96.33	62.92	38.01	65.17	88.27	77.03	88.55	39.70	12.85	48.99
DealMVC [MM'23] [49]	87.50	72.07	87.50	72.28	40.55	72.28	45.53	31.13	45.53	59.55	32.79	61.56
GCFAgg [CVPR'23] [3]	96.67	85.54	96.67	55.06	22.95	61.80	88.27	79.25	88.27	51.51	32.41	61.31
CPSPAN [CVPR'23] [51]	97.83	90.15	97.83	63.30	30.53	63.30	76.26	84.63	85.20	66.83	34.48	**74.12**
SDMVC [TKDE'23] [50]	96.83	86.47	90.00	63.67	43.86	67.79	70.67	83.30	84.92	57.79	33.80	64.57
CVCL [ICCV'23] [46]	95.31	82.36	95.31	58.20	23.73	58.20	56.25	56.01	67.97	61.98	34.68	68.49
MVCAN [CVPR'24] [52]	98.17	91.27	94.59	59.55	42.57	27.18	58.38	66.73	51.58	57.79	35.70	32.24
SCMVC [TMM'24] [47]	97.00	87.11	97.00	73.41	39.76	73.41	93.85	88.44	93.85	50.25	30.70	60.80
SparseMVC (Ours)	**98.33**	**92.01**	**98.33**	**83.15**	**54.62**	**83.15**	**95.25**	**89.86**	**95.25**	**70.10**	**44.90**	70.85

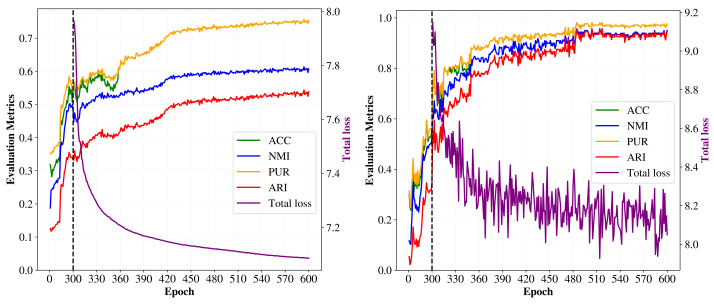
(2) Our method demonstrates stability in clustering performance, as evaluation metrics oscillate upward within a small range and stabilize with increasing epochs, showcasing robust results. In contrast, methods like SURE and DealMVC on ALOI-100 or CVCL and CPSPAN on Wikipedia fail to stabilize, with metrics either degrading significantly after peaking or fluctuating dramatically without consistent improvement. During the early stages of contrastive training when the embedding space's distribution remains uneven, potentially causing abrupt gradient fluctuations. To alleviate the instability, we adopt a dynamic fusion strategy in Sec. 3.2 and a pre-training approach in which only the autoencoder is trained initially.

(3) Although our model is specifically designed to address the challenge: across-view sparsity variations, it also achieves superior performance on dense datasets, such as BRCA, thereby demonstrating its adaptability and broad applicability. This is attributed to the inherent flexibility of our designed sparse autoencoder, which adaptively transitions into a conventional autoencoder when confronted with dense data, thereby prioritizing the reconstruction objective with greater emphasis.

(4) Compared to recent state-of-the-art methods, our approach demonstrates superior feature representation performance by producing clearer boundaries and more compact clusters, as shown in Fig. 5. Its most notable advantage is the ability to disentangle intra-class clusters while preserving inter-class separability, which leads to better scalability and robustness when handling data with large cross-view distribution disparities.

4.3 Convergence Analysis

By analyzing the training curve in Fig. 6, we observe the following key points: (1) The evaluation metrics generally exhibit an oscillatory increase followed by stabilization, with this stability being maintained as the number of training epochs progresses. This observation underscores the model's convergence and its robustness in maintaining stable clustering performance despite optimization challenges. (2) During the early stages of the alignment training phase, the evaluation metrics exhibit a brief dip, which is quickly followed by a recovery and subsequent stabilization at a higher level. The fluctuations observed in the evaluation metrics can be attributed to $\mathcal{L}_{\text{CDA}}$ in Eq. (13), which necessitates the initialization of all network parameters except the autoencoder. (3) Despite significant fluctuations in the loss for small datasets, intriguingly, these fluctuations do not propagate to the evaluation metrics, suggesting that our approach effectively mitigates the influence of instability in the optimization algorithm on the clustering structure.



(a) Out-Scene (b) MSRCV1

Figure 6: Convergence analysis of the training process. The left area of the vertical black dashed line represents the pre-training phase, while the right area stands for the view alignment training process.

Table 4: Clustering results on big (Out-Scene & ALOI-100) and small multi-view datasets.

Methods \ Datasets	Out-Scene			ALOI-100			Wikipedia			MSRCV1		
	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
COMPLETER [CVPR'21] [44]	69.79	55.39	69.79	30.70	62.12	33.63	57.14	53.10	59.31	90.00	87.90	90.00
DCP [TPAMI'22] [45]	56.03	45.59	56.32	34.01	60.28	37.32	45.31	43.16	46.32	25.71	23.25	27.14
MFLVC [CVPR'22] [48]	58.97	51.31	58.97	33.17	73.28	33.17	40.12	27.52	41.70	63.33	66.11	64.29
DSMVC [CVPR'22] [43]	62.13	53.01	64.25	71.52	90.87	72.72	60.32	54.74	62.19	64.29	54.29	64.29
SURE [TPAMI'22] [62]	60.97	48.09	60.97	10.13	34.19	11.90	50.65	39.97	54.11	91.43	85.84	91.43
DealMVC [MM'23] [49]	69.57	59.44	69.57	13.11	48.54	13.10	38.96	37.09	38.96	82.00	75.54	82.00
GCFAgg [CVPR'23] [3]	68.23	57.14	68.23	74.11	88.30	76.63	51.80	45.87	56.57	39.52	31.91	42.86
CPSPAN [CVPR'23] [51]	59.15	50.46	59.15	56.96	78.78	67.99	22.08	8.35	24.39	67.62	69.83	89.52
SDMVC [TKDE'23] [50]	56.03	46.18	59.93	52.02	74.70	56.56	55.99	53.98	62.05	59.52	52.51	45.24
CVCL [ICCV'23] [46]	73.51	59.59	73.51	21.86	43.13	23.29	14.17	42.81	32.69	48.44	84.57	90.62
MVCAN [CVPR'24] [52]	70.98	58.23	49.95	67.48	83.78	56.71	59.02	**55.81**	67.97	71.54	60.19	71.54
SCMVC [TMM'24] [47]	71.54	60.19	71.54	77.72	89.42	81.05	53.54	35.59	55.84	90.95	83.92	90.95
SparseMVC (Ours)	**77.49**	**63.34**	**77.49**	**82.21**	**92.65**	**84.19**	**61.04**	54.79	**62.91**	**97.14**	**94.22**	**97.14**

4.4 Ablation Study

Loss Function To ensure the rigor of ablation experiment, we selected three datasets with significant variations in view sparsity as shown in Fig. 1, ensuring the functionality of the SAA. We assessed the effectiveness of individual losses in the total loss Eq. (14) of SparseMVC, as presented in Table 5.

Specifically, we use $\mathcal{L}_{\text{recon}}$ as the baseline and find that adding either $\mathcal{L}_{\text{entropy}}$ or $\mathcal{L}_{\text{CDA}}$ improves performance. $\mathcal{L}_{\text{entropy}}$ yields substantial improvements on ALOI-100 and Dermatology, which have stronger variations in view sparsity. These improvements highlight the effectiveness of adaptive encoding and cross-view distribution alignment as robust constraints that contribute positively to the overall model training process. Moreover, when all losses are activated simultaneously, the model achieves optimal performance, suggesting that $\mathcal{L}_{\text{entropy}}$ and $\mathcal{L}_{\text{CDA}}$ complement each other synergistically. Notably, their integration does not introduce any mutual interference, further underlining the coherence and compatibility of these objectives in driving superior learning outcomes.

Table 5: Comparison of different loss function combinations.

Datasets	Loss Function			Evaluation Metrics		
	$\mathcal{L}_{\text{recon}}$	$\mathcal{L}_{\text{entropy}}$	$\mathcal{L}_{\text{CDA}}$	ACC	NMI	PUR
ALOI-100	✓			45.27	71.21	30.41
	✓	✓		66.35	81.65	70.62
	✓		✓	64.11	80.23	67.33
	✓	✓	✓	**82.21**	**92.65**	**84.19**
Dermatology	✓			70.11	74.41	59.69
	✓	✓		75.70	83.36	69.36
	✓		✓	70.95	71.06	83.80
	✓	✓	✓	**95.25**	**89.86**	**95.25**
MSRCV1	✓			58.57	48.63	32.76
	✓	✓		70.48	65.27	52.08
	✓		✓	92.38	87.62	92.38
	✓	✓	✓	**97.14**	**94.22**	**97.14**

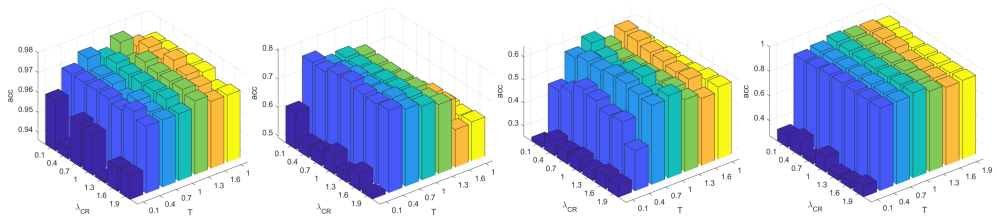
Components Our approach focuses on view-level structural sparsity, specifically the sparsity variation across views within the same multi-view data. This differs from data-level sparsity methods, which typically apply uniform sparse encoding to all views without explicitly considering the heterogeneity of inter-view sparsity. To further validate the effectiveness of the proposed SAA module, we extend the ablation study in Table 5 by introducing two additional comparative settings: **(i)** ***uniformly sparse encoding applied to all views, which mimics methods designed for data-level sparsity***; and **(ii)** ***adaptive encoding tailored to each view*****.** On top of this, ***we also ablate the CSR module***, which reweights the local features during the late fusion stage. Regardless of whether the CSR module is applied, the results in Table 6 show that the proposed SAA, which leverages adaptive autoencoders, remains effective and consistently achieves superior performance. Results further confirm that the effectiveness arises from the synergy between adaptive encoding and sample reweighting, rather than from the use of sparse autoencoders alone. Collectively, these findings confirm that our method is robust to varying sparsity across views and that the SAA and CSR modules function synergistically rather than independently.

Table 6: Ablation study on different components.

Components \ Datasets	ALOI-100			Dermatology			MSRCV1		
	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
all sparse autoencoders w/o CSR	78.56	88.92	81.12	77.37	74.38	86.03	91.90	88.56	91.90
all sparse autoencoders	80.42	89.19	82.44	89.11	78.37	89.11	92.38	88.61	92.38
adaptive autoencoders w/o CSR	81.04	89.58	83.17	88.83	78.23	88.83	95.71	92.69	95.71
adaptive autoencoders (Ours)	**82.21**	**92.65**	**84.19**	**95.25**	**89.86**	**95.25**	**97.14**	**94.22**	**97.14**

4.5 Parameter Sensitivity Analysis

We selected the temperature parameter τ in $\mathcal{L}_{\text{CDA}}$ and the constraint ratio coefficient λ_{CR}, the ratio of $\mathcal{L}_{\text{SAA}}$ to $\mathcal{L}_{\text{CDA}}$, as the two parameters for analysis. Both coefficients were set to gradually increase from 0.1 to 1.9 with a step size of 0.3.



(a) Synthetic3d (b) Out-Scene (c) BRCA (d) MSRCV1

Figure 7: The bar chart of clustering accuracy varying with different values of τ and λ_{CR}.

In Fig. 7, we can discover that the accuracy initially increases and then decreases as τ increases, remaining relatively stable within a range around 1.0. The influence of λ_{CR} on clustering performance is comparatively minor, with a negligible impact when τ is within the range of 0.4 to 1.0. In light of Sec. 3.3, $\mathcal{L}_{\text{CDA}}$ incorporates the smoothing property of the logarithmic function, which diminishes the direct effect of λ_{CR} adjustments on the gradient. In contrast, changes to τ significantly influence clusters separability and alignment performance by modulating the nonlinear response of the softmax function. A smaller τ enhances class separability, while a larger τ emphasizes global consistency. Therefore, we set the default values of τ and λ_{CR} to 1.0 in the loss function.

Our method is largely insensitive to hyperparameter changes. To begin with, significant performance degradation only occurs when the temperature coefficient $\tau \leq 0.4$ or ≥ 1.6. The extremely small value of 0.1 is chosen to probe the lower bound of performance degradation and is rarely used in practical applications. Furthermore, the performance fluctuation mainly occurs along the τ-axis, whereas it remains relatively insensitive to changes in the constraint ratio coefficient λ_{CR}. In practice, it is the relative weighting between loss terms that is more commonly adjusted. In addition, the accuracy-axis was intentionally truncated to better highlight the differences, accentuating the visual disparity. Finally, regarding the concern that hyperparameters are not easy to tune in practice, our method maintains stable performance even under noticeable loss fluctuations, as illustrated in Fig. 6.

5 Conclusion

This paper highlights a frequently overlooked issue in deep multi-view learning: varying sparsity ratios across views. Therefore, we systematically define, quantify, and analyze cross-view sparsity variation as a fundamental characteristic of multi-view data. Our entire framework, SparseMVC, is designed to handle view-level sparsity variations with a complete data-driven and tightly integrated architecture. To tackle sparsity variation, we propose an adaptive encoding strategy that uses the sparsity ratio of each view as prior knowledge, enabling the encoder to switch between standard and sparse forms with appropriate constraint strengths. Additionally, we introduce a series of interdependent mechanisms to mitigate the side effects of representational divergence caused by non-uniform encoding. Specifically, a correlation-guided fusion strategy leverages global-to-local feature relationships from the early stages to guide the weighting of local features in late fusion. Moreover, a distribution alignment module structurally constrains the fused representations, enhancing cross-view complementarity in the final stage. Comprehensive experiments and detailed dissections of each module validate the efficacy of SparseMVC. We hope this work inspires greater attention to the intrinsic characteristics of data and to the design of architectures driven by data.

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China under grants 62522604 and 62476258, the Natural Science Foundation of Hubei Province under grant 2025AFA113, the Key Laboratory of Target Cognition and Application Technology (No. 2023-CXPT-LC-005), and the project was supported by the Fundamental Research Funds for National Universities, China University of Geosciences (Wuhan), No.2024XLB7.

References

[1] Siwei Wang, Xinwang Liu, Suyuan Liu, Jiaqi Jin, Wenxuan Tu, Xinzhong Zhu, and En Zhu. Align then fusion: Generalized large-scale multi-view clustering with anchor matching correspondences. *Advances in Neural Information Processing Systems*, 35:5882–5895, 2022.

[2] Xin Zou, Chang Tang, Xiao Zheng, Kun Sun, Wei Zhang, and Deqiong Ding. Inclusivity induced adaptive graph learning for multi-view clustering. *Knowledge-Based Systems*, 267:110424, 2023.

[3] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19863–19872, 2023.

[4] Jing Wang and Songhe Feng. Contrastive and view-interaction structure learning for multi-view clustering. In *International Joint Conference on Artificial Intelligence*, pages 5055–5063, 2024.

[5] Yuang Xiao, Dong Yang, Jiaxin Li, Xin Zou, Hua Zhou, and Chang Tang. Dual alignment feature embedding network for multi-omics data clustering. *Knowledge-Based Systems*, 309:112774, 2025.

[6] Shiping Wang, Zhaoliang Chen, Shide Du, and Zhouchen Lin. Learning deep sparse regularizers with applications to multi-view clustering and semi-supervised classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5042–5055, 2021.

[7] Zhenglai Li, Chang Tang, Xinwang Liu, Xiao Zheng, Wei Zhang, and En Zhu. Consensus graph learning for multi-view clustering. *IEEE Transactions on Multimedia*, 24:2461–2472, 2021.

[8] Chao Zhang, Huaxiong Li, Wei Lv, Zizheng Huang, Yang Gao, and Chunlin Chen. Enhanced tensor low-rank and sparse representation recovery for incomplete multi-view clustering. In *AAAI Conference on Artificial Intelligence*, volume 37, pages 11174–11182, 2023.

[9] Pan Xiao, Peijie Qiu, Sungmin Ha, Abdalla Bani, Shuang Zhou, and Aristeidis Sotiras. Sc-vae: Sparse coding-based variational autoencoder with learned ista. *Pattern Recognition*, page 111187, 2024.

[10] Canyi Lu, Shuicheng Yan, and Zhouchen Lin. Convex sparse spectral clustering: Single-view to multi-view. *IEEE Transactions on Image Processing*, 25(6):2833–2843, 2016.

[11] Chengliang Liu, Zhihao Wu, Jie Wen, Yong Xu, and Chao Huang. Localized sparse incomplete multi-view clustering. *IEEE Transactions on Multimedia*, 25:5539–5551, 2022.

[12] Yik Lung Pang, Changjae Oh, and Andrea Cavallaro. Sparse multi-view hand-object reconstruction for unseen environments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–810, 2024.

[13] Kaiyi Xu, Minhui Wang, Xin Zou, Chengfu Ji, Hua Zhou, and Chang Tang. Cancer–drug response prediction via feature aggregation and association graph learning. *Engineering Applications of Artificial Intelligence*, 151:110671, 2025.

[14] Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase "highlighted tokens" in mllms: Revisiting visual holistic context retention. *arXiv preprint arXiv:2510.02912*, 2025.

[15] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[16] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279, 2010.

[17] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3565, 2014.

[18] Timothy Danford, Alex Rolfe, and David Gifford. Gse: a comprehensive database system for the representation, retrieval, and analysis of microarray data. In *Biocomputing*, pages 539–550. World Scientific, 2008.

[19] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[20] 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.

[21] Youwei Liang, Dong Huang, Chang-Dong Wang, and Philip S Yu. Multi-view graph learning by joint modeling of consistency and inconsistency. *IEEE transactions on neural networks and learning systems*, 35(2):2848–2862, 2022.

[22] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

[23] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *International Conference on Machine Learning*, pages 129–136, 2009.

[24] Meng Qu, Jian Tang, Jingbo Shang, Xiang Ren, Ming Zhang, and Jiawei Han. An attention-based collaboration framework for multi-view network representation learning. In *ACM on Conference on Information and Knowledge Management*, pages 1767–1776, 2017.

[25] Kaixuan Yao, Jiye Liang, Jianqing Liang, Ming Li, and Feilong Cao. Multi-view graph convolutional networks with attention mechanism. *Artificial Intelligence*, 307:103708, 2022.

[26] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30:1771–1783, 2021.

[27] Jun Wang, Chang Tang, Zhiguo Wan, Wei Zhang, Kun Sun, and Albert Y Zomaya. Efficient and effective one-step multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):12224–12235, 2023.

[28] Shudong Huang, Zhao Kang, Ivor W Tsang, and Zenglin Xu. Auto-weighted multi-view clustering via kernelized graph learning. *Pattern Recognition*, 88:174–184, 2019.

[29] Jun Wang, Zhenglai Li, Chang Tang, Suyuan Liu, Xinhang Wan, and Xinwang Liu. Multiple kernel clustering with adaptive multi-scale partition selection. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6641–6652, 2024.

[30] Juncheng Lv, Zhao Kang, Boyu Wang, Luping Ji, and Zenglin Xu. Multi-view subspace clustering via partition fusion. *Information Sciences*, 560:410–423, 2021.

[31] Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20707–20717, 2022.

[32] Xin Zou, Chang Tang, Xiao Zheng, Zhenglai Li, Xiao He, Shan An, and Xinwang Liu. Dpnet: Dynamic poly-attention network for trustworthy multi-modal classification. In *Proceedings of the 31st ACM international conference on multimedia*, pages 3550–3559, 2023.

[33] Xin Zou, Chang Tang, Wei Zhang, Kun Sun, and Liangxiao Jiang. Hierarchical attention learning for multimodal classification. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 936–941. IEEE, 2023.

[34] Jian Zhu, Xin Zou, Lei Liu, Zhangmin Huang, Ying Zhang, Chang Tang, and Li-Rong Dai. Trusted mamba contrastive network for multi-view clustering. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[35] Xiao Yu, Hui Liu, Yuxiu Lin, Nan Liu, and Shanbao Sun. Sample-level weights learning for multi-view clustering on spectral rotation. *Information Sciences*, 619:38–51, 2023.

[36] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997.

[37] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1794–1801. IEEE, 2009.

[38] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in Neural Information Processing Systems*, 6, 1993.

[39] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[40] Andrew Ng et al. Sparse autoencoder. *CS294A Lecture Notes*, 72(2011):1–19, 2011.

[41] Xi Peng, Jiashi Feng, Shijie Xiao, Wei-Yun Yau, Joey Tianyi Zhou, and Songfan Yang. Structured autoencoders for subspace clustering. *IEEE Transactions on Image Processing*, 27(10):5076–5086, 2018.

[42] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

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

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

[45] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2022.

[46] Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *IEEE/CVF International Conference on Computer Vision*, pages 16752–16761, 2023.

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

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

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

[50] Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, S Yu Philip, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7470–7482, 2023.

[51] Jiaqi Jin, Siwei Wang, Zhibin Dong, Xinwang Liu, and En Zhu. Deep incomplete multi-view clustering with cross-view partial sample and prototype alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11600–11609, 2023.

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

[53] John Winn and Nebojsa Jojic. Locus: Learning object classes with unsupervised segmentation. In *International Conference on Computer Vision*, volume 1, pages 756–763. IEEE, 2005.

[54] H Altay Güvenir, Gülşen Demiröz, and Nilsel Ilter. Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine*, 13(3):147–165, 1998.

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

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

[57] Cancer Genome Atlas Research Network. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498, 2015.

[58] Daniel C Koboldt, Robert S Fulton, Michael D McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua F McMichael, Lucinda L Fulton, David J Dooling, Li Ding, Elaine R Mardis, et al. Tcga-network, comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[59] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. *Advances in Neural Information Processing Systems*, 24, 2011.

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

[61] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[62] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1055–1069, 2022.

[63] Geping Yang, Shusen Yang, Yiyang Yang, Xiang Chen, Can Chen, Zhiguo Gong, and Zhifeng Hao. Spgmvc: Multiview clustering via partitioning the signed prototype graph. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[64] Kaixuan Yao, Jiye Liang, Jianqing Liang, Ming Li, and Feilong Cao. Multi-view graph convolutional networks with attention mechanism. *Artificial Intelligence*, 307:103708, 2022.

[65] Fangdi Wang, Jiaqi Jin, Jingtao Hu, Suyuan Liu, Xihong Yang, Siwei Wang, Xinwang Liu, and En Zhu. Evaluate then cooperate: Shapley-based view cooperation enhancement for multi-view clustering. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[66] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

[67] Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *Forty-second International Conference on Machine Learning (ICML)*, 2025.

[68] Yu Kang, Erwei Liu, Kaichi Zou, Xiuyun Wang, and Huaqing Zhang. Sparse clustering algorithm based on multi-domain dimensionality reduction autoencoder. *Mathematics*, 12(10):1526, 2024.

[69] Junwei Han, Jinglin Xu, Feiping Nie, and Xuelong Li. Multi-view k-means clustering with adaptive sparse memberships and weight allocation. *IEEE Transactions on Knowledge and Data Engineering*, 34(2):816–827, 2020.

[70] Qiyue Yin, Shu Wu, Ran He, and Liang Wang. Multi-view clustering via pairwise sparse subspace representation. *Neurocomputing*, 156:12–21, 2015.

[71] Zhanxuan Hu, Feiping Nie, Wei Chang, Shuzheng Hao, Rong Wang, and Xuelong Li. Multi-view spectral clustering via sparse graph learning. *Neurocomputing*, 384:1–10, 2020.

[72] Jie Chen, Shengxiang Yang, Xi Peng, Dezhong Peng, and Zhu Wang. Augmented sparse representation for incomplete multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3):4058–4071, 2022.

[73] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[74] Tianyu Hua, Wenxiao Wang, Zihui Xue, Sucheng Ren, Yue Wang, and Hang Zhao. On feature decorrelation in self-supervised learning. In *IEEE/CVF International Conference on Computer Vision*, pages 9598–9608, 2021.

[75] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013.

[76] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via $\ell 1$ minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[77] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly articulate the core problem addressed by the paper, cross-view sparsity variations in multi-view clustering, and emphasize the limitations of existing methods in handling this structural heterogeneity. The proposed solution, SparseMVC, is concisely described as incorporating sparsity-aware constraints and a view-adaptive autoencoder design. These claims are fully supported by both the theoretical formulation and the experimental results across diverse datasets, including those with extreme sparsity imbalance.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Appendix F.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Please refer to Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please refer to Section 4.1 for details on reproducing the main experimental results presented in the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The data and code of the proposed algorithm will be uploaded in a zip file along with the supplementary material. All baseline methods and datasets evaluated in the paper are publicly available and can be reproduced by following the comparison protocol outlined in Section 4.1.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Sections 4.1 and 4.5 for details on the experimental settings and hyperparameter configurations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper does not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Sections 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have carefully reviewed the NeurIPS Code of Ethics and confirm that our research fully complies with its principles.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In the current field of artificial intelligence, there is an abundance of research focused on meticulously refining models and methods. This paper serves as a reminder to shift the focus toward more in-depth analysis of the inherent issues within the data itself, encouraging the design of networks based on data characteristics.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please refer to Section 4.1. All baselines and datasets employed in this paper are appropriately cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

For completeness, we include additional discussions and experimental details in the appendices. A summary of the contents is listed below:

- **Appendix A** Algorithm and Comparison with Prior Work.
- **Appendix B** Further Experiments.
- **Appendix C** Visualization.
- **Appendix D** Computational Complexity.
- **Appendix E** Theoretical Analysis.
- **Appendix F** Limitations and Future Work.

A Algorithm and Comparison with Prior Work

A.1 Algorithm

The training procedure for SparseMVC is described in Algorithm (1).

Algorithm 1 Training Steps for SparseMVC

Input: Multi-view data $\{\boldsymbol{X}_v\}_{v=1}^V$, cluster number K, and number of training epochs E_{pre}, E_{con}.
Output: Late-stage fusion representation $\overline{\boldsymbol{Y}}$.

```
Initialize random seed and select Adam optimizer.
for epoch = 1 : E_pre + E_con do
    Update {Z_v}_{v=1}^V by minimizing {L_recon^v}_{v=1}^V and {L_entropy^v}_{v=1}^V utilizing Eqs. (2) and (4).
    Update Z̄, formed by the concatenation of {Z_v}_{v=1}^V, utilizing Eq. (2) and Eq. (4).
    if epoch > E_pre then
        Update weights {W_v^i}_{v=1}^V by Eq. (9).
        Update Ȳ by minimizing L_CDA utilizing Eq. (13).
    end if
end for
Perform K-means clustering on representation Ȳ.
```

A.2 Comparison with Prior Work

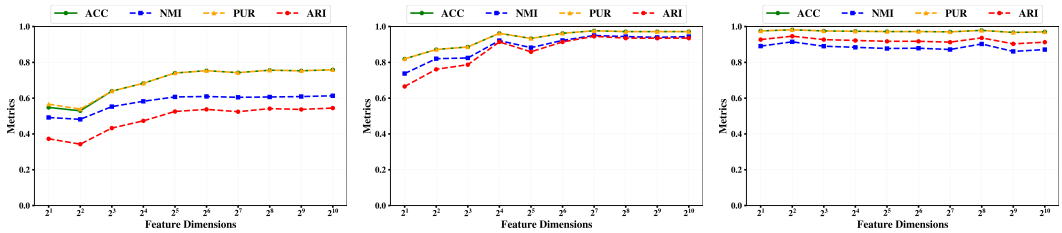
To contextualize our contributions, we present a comparative discussion with representative dynamic weighting methods proposed in the literature. Dynamic weighting has been widely explored in multi-view learning, primarily through attention-based fusion mechanisms or optimization-driven strategies. Methods such as GCFAgg [3] employ attention to emphasize discrepancies among local views, aiming to refine feature alignment. Other approaches, including SPGMVC [63] and MAGCN [64], enhance feature encoding or perform view-level aggregation to improve representational quality. In parallel, techniques such as SCMVC [47], SCE-MVC [65] and TMC [66] utilize mutual information [67] or probabilistic priors to adaptively assign importance to views or losses. Furthermore, recent efforts have introduced view-invariant representations [46] and prototype-guided learning [52] to mitigate cross-view variability. Despite these advances, existing methods often overlook the inherent inconsistencies of sparsity between views, a phenomenon that can severely degrade the effectiveness of fusion [8, 9, 68]. In contrast, our proposed approach is motivated by the need to explicitly characterize and adapt to such cross-view sparsity variations. Specifically, our framework preserves global features obtained during early fusion and integrates both global and local view representations into the fusion process. The core of our design, the correlation-informed sample reweighting module, dynamically adjusts fusion weights based on the learned correlation between global and local views, thereby enabling fine-grained, sample-specific adaptation.

While MVASM [69] addresses the challenges of ambiguous class assignments in multi-view data by introducing an adaptive sparse membership matrix, our method introduces adaptive autoencoders with view-specific encoding strategies. Moreover, unlike existing techniques [70, 71, 72] that generally apply static or view-level weighting schemes, our method performs late-stage fusion reweighting conditioned on earlier global-local interactions. This design not only enhances fusion fidelity but also

elevates sparsity modeling from the feature level to the view level, offering a principled solution to a problem rarely addressed in the literature.

B Further Experiments

B.1 Dimension and Feature Compression



(a) Out-Scene (b) MSRCV1 (c) Synthetic3d

Figure 8: The clustering evaluation metric curves with respect to feature ($\boldsymbol{Z}_v$) dimension variations.

We tested the dimensions of view-specific features $\boldsymbol{Z}_v$ extracted by the autoencoder, varying from 2^1 to 2^{10} in Fig 8. The experimental results demonstrated that as the feature dimension increased, all evaluation metrics (ACC, NMI, PUR and ARI) initially exhibited strong oscillations but progressively improved, eventually stabilizing within a small and well-performing range.

From the perspective of information-theoretic [73], low-dimensional features have limited encoding capacity that hinders capturing complex data patterns, whereas high-dimensional features offer greater capacity but risk dimensional collapse [74]. The feature compression layers that we designed compress and refine $\boldsymbol{Z}_v$ into $\boldsymbol{Y}_v$, mitigating the risks of overfitting and performance degradation often seen in contrastive learning. Accordingly, the model can avoid the negative effects of excessive dimensionality while leveraging its increased capacity for effective feature extraction. Ultimately, we selected 64 as the dimensionality of $\boldsymbol{Z}_v$ to balance computation and feature capacity.

B.2 Selection of Scaling Factor

Table 7: Ablation study on scaling factor selection.

The external scaling factor $f(s_v)$	the internal ratio coefficient ρ	MSRCV1 **Mean accuracy** [Max accuracy]	ALOI-100 **Mean accuracy** [Max accuracy]
0 or 1	1-$f(s_v)$	**94.29** [95.71]	**77.75** [80.39]
$f(s_v)$	1-$f(s_v)$	**95.24** [95.24]	**79.01** [80.08]
0 or 1	0.05	**95.71** [96.67]	**80.82** [81.51]
$f(s_v)$	0.05	**97.14** [97.62]	**82.21** [82.93]

To investigate the relative importance of the external scaling factor $f(s_v)$ and the internal ratio coefficient ρ in modulating sparsity within the sparse autoencoder, we conducted a set of controlled experiments in which one parameter was kept constant while systematically varying the other. This decoupled analysis enables a clearer understanding of how each component contributes to the overall behavior of the model. Empirical results, as shown in Table 7, demonstrate that allowing $f(s_v)$ to be adaptively optimized while fixing ρ at a reasonable constant leads to superior performance in multiple evaluation metrics. The smooth variation of the constraint strength is better than directly switching between zero and one, which represents whether the sparse constraint is applied. In contrast, varying ρ while keeping $f(s_v)$ static yields relatively suboptimal results. These findings suggest that external view-level scaling plays a more critical role in capturing cross-view sparsity dynamics, highlighting the effectiveness of our adaptive design.

B.3 The large-scale datasets

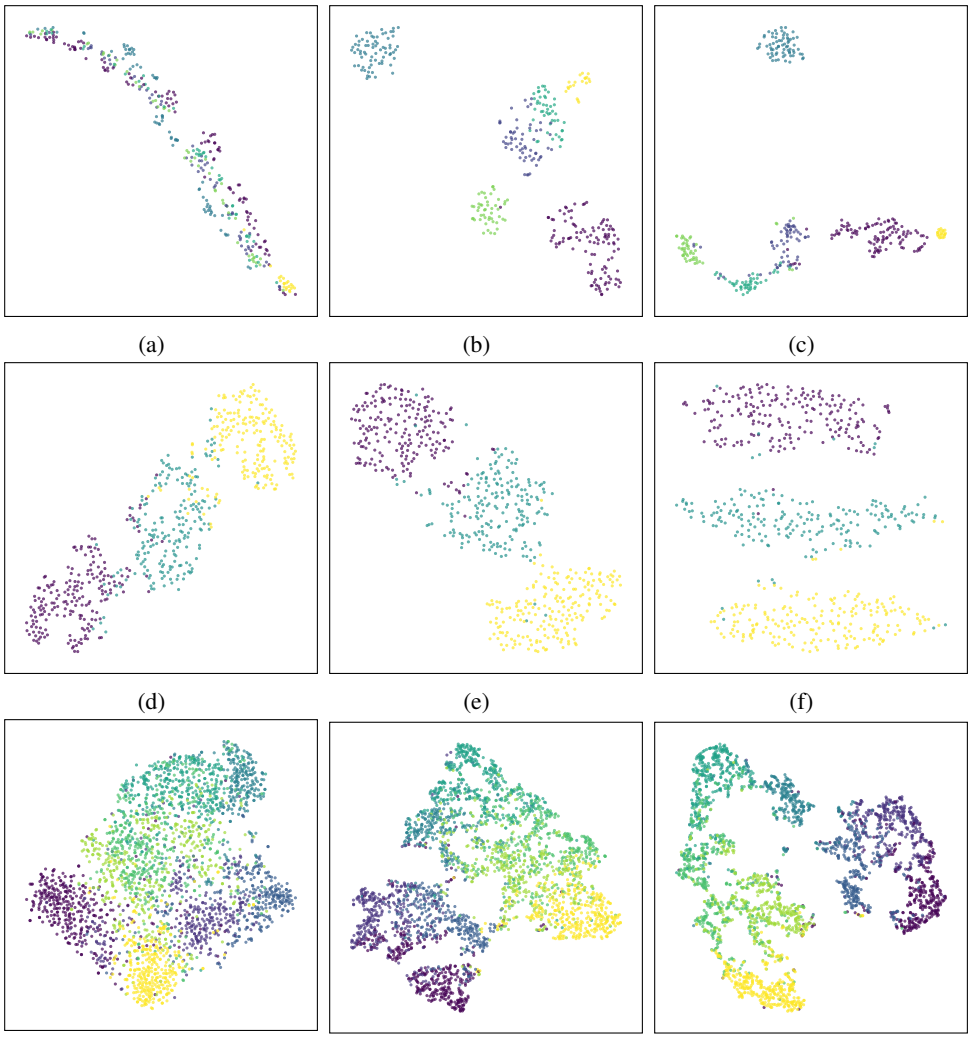
To evaluate the effectiveness and generalization ability of our proposed method, SparseMVC, we conduct additional experiments on large-scale datasets comprising over 8,000 samples, specifically GSE [18] and Animal [75]. The GSE dataset encompasses multi-omics data across 27 categories,

Table 8: Clustering results on the large-scale multi-view datasets.

Datasets View Sparsity Ratio Methods	GSE (8,200 samples) [0.877, 0.005] ACC	NMI	PUR	Animal (11,673 samples) [0.589, 0.179, 0.355, 0.467] ACC	NMI	PUR
MVCAN [CVPR'24]	71.40	75.41	72.29	12.06	10.25	15.48
SCMVC [TMM'24]	62.80	74.13	65.72	16.96	15.05	20.15
SparseMVC (Ours)	**73.18**	**78.42**	**74.26**	**19.90**	**17.92**	**24.13**

capturing a broad range of biological conditions, and is extensively utilized in bioinformatics research. Meanwhile, the Animal dataset consists of image data featuring animals with diverse attributes, spanning 20 categories. The detailed dataset information and comparison results, presented in Table 8, highlight the performance of our approach against the latest two methods employed in our manuscript. Clustering results on large-scale multi-view datasets further emphasize the superiority of our method, demonstrating its promising applicability in other data environments and scenarios.

C Visualization



(a) (b) (c) (d) (e) (f) (g) $\widehat{\boldsymbol{X}}_v$ features (h) $\overline{\boldsymbol{Z}}$ features (i) $\overline{\boldsymbol{Y}}$ features

Figure 9: The t-SNE visualizations of the original reconstruction features ($\widehat{\boldsymbol{X}}_v$), the early fusion features ($\overline{\boldsymbol{Z}}$), and the late fusion features ($\overline{\boldsymbol{Y}}$) on datasets: Dermatology (a-c), Synthetic3d (d-f), and Out-Scene (g-i).

Figure 9 uses t-SNE for dimensionality reduction, mapping high-dimensional feature vectors into a two-dimensional space to facilitate a clearer and more insightful visualization of the distribution and structure of the data. The visualization reveals that the original reconstruction data, $\widehat{\boldsymbol{X}}_v$, exhibit substantial interclass entanglement, with considerable overlap between different categories. In contrast, the preliminary global features, $\boldsymbol{Z}_v$, derived from autoencoder pretraining, show a notable reduction in this entanglement, suggesting an early phase of disentanglement compared to the raw features. Further refinement through multi-view alignment results in the final fused features, $\boldsymbol{Y}_{\text{global}}$, which exhibit the most pronounced disentanglement, yielding well-defined, separate clusters. These findings are further corroborated by the results in ablation analysis, which demonstrate that $\boldsymbol{Y}_{\text{global}}$ consistently outperforms $\boldsymbol{Z}_v$ in clustering tasks, particularly in distance-based methods like K-means. This performance enhancement underscores the pivotal role of the multi-view alignment and fusion process in improving feature separability and discriminative capacity, ultimately leading to more accurate and meaningful clustering results.

D Computational Complexity

The total loss function, as defined in Eq. (14), includes two main terms: $\mathcal{L}_{\text{SAA}}$ and $\mathcal{L}_{\text{CVDA}}$. The term $\mathcal{L}_{\text{SAA}}$ involves the reconstruction loss and the entropy-matching loss, both computed for each of the V views, each containing N samples. Since the feature dimension F is constant and does not affect the computational scaling, the complexity of this term is determined by the number of views and samples, resulting in $O(VN)$. The term $\mathcal{L}_{\text{CVDA}}$ requires pairwise similarity computations between N samples. For each pair of samples, the dot product computation has a constant cost of $O(1)$, leading to a total complexity of $O(N^2)$ for computing all pairwise similarities. The contrastive loss further requires computing the numerator, denominator, and logarithmic terms for each sample, which does not increase the overall scaling beyond $O(N^2)$. When combining both terms, the reconstruction and entropy-matching losses contribute $O(VN)$, while the pairwise similarity computations dominate with $O(N^2)$ when N is sufficiently large. Thus, the overall computational complexity of the total loss function is $O(N^2)$.

E Theoretical Analysis

When learning from heterogeneous multi-view data, structural disparities, especially in sparsity patterns, present significant challenges to unified representation learning. Differences in modality, sampling granularity, and view incompleteness lead to varying information densities across views, making fixed encoder architectures and uniform loss formulations inherently suboptimal. To address this issue, we propose the Sparse Autoencoder with Adaptive Constraints (SAA), which dynamically adjusts the strength of sparsity constraints based on sparsity ratio s_v of each view. This mechanism enables the encoder to balance compression and expressiveness in a view-aware manner, thereby facilitating the alignment of latent representations across structurally diverse views. In what follows, we provide a preliminary theoretical foundation for SAA, organized around a set of key questions that clarify its motivation, theoretical grounding, and coding formulation:

❶ **Why** should the sparsity constraint be adaptive?
When dealing with views of varying characteristics, uniform treatment of data may result in suboptimal representations, particularly when the model overly compresses informative in dense views or overemphasizes noisy in sparse ones. Accordingly, by the principle of minimal redundancy maximum relevance (MRMR), we can establish trade-offs between different views, balancing the fidelity of reconstruction and the complexity of representation.

❷ **What** exactly are we coding or preserving?
Grounded in compressed sensing and optimally sparse representations [76, 77], we encode the essential structure of sparse inputs, which, though embedded in high-dimensional space, intrinsically conform to low-dimensional semantic subspaces. Therefore, sparse activation suffices to capture their core features, with sparse coding providing a stable, efficient, and redundancy-minimizing representation aligned with this inherent geometry.

E.1 Why should sparsity constraints be adaptive?

MRMR Principle in Unsupervised Representation Learning The minimal redundancy maximum relevance (MRMR) principle aims to learn a set of features that preserves as much information about the input as possible, while minimizing redundancy among the features. In the context of multi-view autoencoders, let the hidden representation of view v be $\mathbf{H}^v = \{h_1^v, \ldots, h_H^v\}$. We formalize the MRMR objective as follows:

$$\max_{\mathbf{H}^v} \Big[\underbrace{I(\mathbf{H}^v; \mathbf{X}^v)}_{\text{Max–Relevance}} - \underbrace{\beta \sum_{i \neq j} I(h_i^v; h_j^v)}_{\text{Min–Redundancy}} \Big], \tag{15}$$

where $I(\cdot;\cdot)$ denotes mutual information. The first term encourages the hidden representation to capture the core information of the input data, while the second term penalizes redundancy among the hidden units. The parameter $\beta > 0$ balances relevance and redundancy.

Equivalence of Max–Relevance and Reconstruction Loss An autoencoder minimizes the mean squared reconstruction error, which can be shown to be equivalent to maximizing the mutual information between its code and the input, up to an additive constant. Specifically, the following chain of equalities can be written:

$$\begin{aligned} \mathcal{L}_{\text{recon}}^v &= \frac{1}{N} \sum_{j=1}^{N} \|\hat{\mathbf{x}}_j^v - \mathbf{x}_j^v\|^2 \\ &\approx -\mathbb{E}_{\mathbf{x} \sim \mathbf{X}^v} \left[\log p_\theta(\mathbf{x} \mid g_\phi(\mathbf{x})) \right] + C \\ &\approx -I(\mathbf{H}^v; \mathbf{X}^v) + C, \end{aligned} \tag{16}$$

where the first approximation uses a Gaussian likelihood with fixed variance to relate MSE to log-likelihood, and the second uses the identity $I(\mathbf{H}^v; \mathbf{X}^v) = H(\mathbf{X}^v) + \mathbb{E}[\log p_\theta(\mathbf{x} \mid \mathbf{h})] - C$. Thus, minimizing $\mathcal{L}_{\text{recon}}^v$ is equivalent to maximizing $I(\mathbf{H}^v; \mathbf{X}^v)$, realizing the MRMR max–relevance criterion. Equation (16) captures the full derivation in one display.

Equivalence of Min–Redundancy and Sparse KL Divergence Starting from the sparsity constraint term based on the entropy-matching loss in (15):

$$\begin{aligned} \sum_{i \neq j} I(h_i^v; h_j^v) &\approx \sum_{k=1}^{H} D_{\text{KL}}(\rho \| \hat{h}_k^v) \\ &= \sum_{k=1}^{H} \Big[\rho \ln \frac{\rho}{\hat{h}_k^v} + (1-\rho) \ln \frac{1-\rho}{1-\hat{h}_k^v} \Big] \\ &\approx -\sum_{k=1}^{H} H\big(h_k^v\big) + C(\rho), \end{aligned} \tag{17}$$

where each hidden activation $\hat{h}_k^v$ is treated as $\text{Bernoulli}(\rho)$, and $D_{\text{KL}}(\rho \| \hat{h}_k^v)$ is the KL divergence as defined in [40]. Thus, minimizing the sparse KL divergence is equivalent to minimizing redundancy among hidden units, according to the minimal redundancy objective of MRMR.

Derivation of the Adaptive Sparse Autoencoder Loss Substituting (2) and (4) into the MRMR formulation (15), and setting $\beta = f(s_v)$ to adapt to the sparsity s_v of view v, we obtain:

$$\begin{aligned} \min_{\mathbf{H}^v} \Big[I(\mathbf{H}^v; \mathbf{X}^v) - f(s_v) \sum_{i \neq j} I(h_i^v; h_j^v) \Big] &\approx \min_{\mathbf{H}^v} \Big[\mathcal{L}_{\text{recon}}^v + f(s_v) \sum_{k=1}^{H} D_{\text{KL}}(\rho \parallel \hat{h}_k^v) \Big] \\ &= \min_{\mathbf{H}^v} \Big[\mathcal{L}_{\text{recon}}^v + \mathcal{L}_{\text{sparse}}^v \Big], \end{aligned} \tag{18}$$

where by definition:

$$\mathcal{L}_{\text{sparse}}^v = f(s_v) \sum_{k=1}^{H} \Big[\rho \ln \frac{\rho}{\hat{h}_k^v} + (1-\rho) \ln \frac{1-\rho}{1-\hat{h}_k^v} \Big], \tag{19}$$

and consequently:

$$\mathcal{L}_{\text{SAA}} = \sum_{v=1}^{V} \Big(\mathcal{L}_{\text{recon}}^{v} + \mathcal{L}_{\text{sparse}}^{v} \Big). \tag{20}$$

Conclusion Eqs. (2) and (19) concretely instantiate the MRMR criteria of *max–relevance* and *min–redundancy* within the SparseMVC framework. By setting the redundancy coefficient $\beta = f(s_v)$, the penalty based on KL divergence for each view is automatically scaled according to its measured sparsity s_v, thereby establishing a principled, data-driven balance between information preservation and redundancy reduction. This adaptive approach is both logical and essential: views with low intrinsic sparsity may already encode compressed information, requiring less stringent sparsity enforcement, while denser views, which often contain redundant or noisy components, benefit from more robust regularization. Thus, adaptively modulating the sparsity constraint ensures an optimal trade-off between retaining critical information and minimizing redundancy.

E.2 What exactly are we coding or preserving?

This question is essentially an analysis of the rationality and effectiveness of using sparsity constraint and sparse autoencoder to represent sparse data. Sparse data is typically characterized by low information density, where only the nonzero entries convey structural information. This observation leads naturally to the principle that representations of sparse inputs should also be sparse, activating only the minimal subset of neurons corresponding to the nonzero dimensions, thereby aligning the sparsity of the input with the representational sparsity. Why does sparse coding offer a more principled and effective encoding than conventional dense mappings for data riddled with zero-valued? In the following, we address this question from the perspective of compressed sensing and optimally sparse representations.

Sparse Input Decomposition Let the j-th sample in view v be denoted as:

$$x_j^v \in \mathbb{R}^{n^v}, \quad \|x_j^v\|_0 = k \ll n^v, \tag{21}$$

with support $\Omega = \{i \mid x_{j,i}^v \neq 0\}$. We decompose x_j^v as follows:

$$x_j^v = x_{j,\Omega}^v + x_{j,\Omega^c}^v, \quad x_{j,\Omega^c}^v = 0. \tag{22}$$

In sparse coding, the goal is to represent sparse inputs accurately by learning a compact and informative latent representation. The reconstruction error represents the difference between the original input and its approximation from the learned representation. In the case of sparse inputs, the information is mainly contained in the non-zero elements, while the zero elements carry no information. Therefore, it is important to split the reconstruction error into components corresponding to the non-zero and zero elements. This allows for more efficient learning by focusing only on the non-zero components.

Let the dictionary $D^v \in \mathbb{R}^{n^v \times m^v}$ represent the basis for the sparse representations and let α_j^v be the coding feature for the j-th sample in view v. The original input x_j^v can be decomposed into two parts based on its support Ω, which is the set of indices where the input is non-zero, and its complement Ω^c, where the input is zero. This decomposition allows us to calculate the reconstruction error for each part separately.

Reconstruction Error Splitting The reconstruction error can be split into two parts, one for the non-zero components (support set Ω) and one for the zero components (complement set Ω^c):

$$\begin{aligned} \|x_j^v - D^v \alpha_j^v\|_2^2 &= \|x_{j,\Omega}^v - D_\Omega^v \alpha_j^v\|_2^2 + \|x_{j,\Omega^c}^v - D_{\Omega^c}^v \alpha_j^v\|_2^2 \\ &= \|x_{j,\Omega}^v - D_\Omega^v \alpha_j^v\|_2^2 + \|D_{\Omega^c}^v \alpha_j^v\|_2^2, \end{aligned} \tag{23}$$

In the above equation, the first term represents the error due to the non-zero part of the input (corresponding to the support set Ω), and the second term represents the error for the zero components (the complement set Ω^c). The second term is expected to be small, as we assume that the input has only a few non-zero entries.

Support Restriction To enforce the sparsity of the representation, we impose a restriction on the model such that the reconstruction error from the zero components is forced to be zero. This can be achieved by setting the corresponding coefficients for the zero entries in the input to be zero as well:

$$D^v_{\Omega^c}\,\alpha^v_j = 0$$

This ensures that the model does not fit the zero parts of the input, effectively restricting the fitting process to the non-zero support:

$$\|x^v_j - D^v\alpha^v_j\|_2^2 = \|x^v_{j,\Omega} - D^v_\Omega\alpha^v_j\|_2^2. \tag{24}$$

By enforcing this constraint, we ensure the model's effort is focused only on learning the most informative parts of the input, which leads to a more efficient and meaningful representation. This is particularly useful in scenarios where inputs are sparse, ensuring that the learned representation is both stable and sparse, capturing the essential structure of the data.

Sparse Coding and Recoverability Our focus is on probing the sparse coding problem:

$$\min_{\|\alpha\|_0\le s}\left\|x^v_{j,\Omega} - D^v_\Omega\alpha\right\|_2^2, \tag{25}$$

where $s \ll m^v$. *Compressed sensing theory* asserts that when D^v_Ω satisfies the Restricted Isometry Property, any sufficiently sparse vector can be stably recovered [77]. Additionally, *optimally sparse representation*, as defined by Donoho and Elad [76], guarantees that if:

$$s < \frac{1}{2}\,\mathrm{spark}\big(D^v_\Omega\big), \tag{26}$$

then the solution

$$\alpha^{v*}_j = \arg\min_{\|\alpha\|_0\le s}\left\|x^v_{j,\Omega} - D^v_\Omega\alpha\right\|_2^2, \tag{27}$$

is both unique and recoverable. This guarantee underpins the choice of sparse coding for inputs that are themselves sparse, ensuring fidelity and stability.

Implementation via Sparse Autoencoder In a standard autoencoder, the encoder generates the following activation:

$$h^v_j = \sigma\big(W^v x^v_j + b^v\big). \tag{28}$$

In conventional dense encoding, each sample—regardless of the fact that only a small fraction of its dimensions are nonzero—is mapped to a fully dense hidden representation. Consequently, the encoder is forced to infer missing dimensions, leading to the creation of spurious relationships that misrepresent the true structure of the data. In contrast, sparse inputs lie within a low-dimensional subspace or manifold embedded in the high-dimensional ambient space. To address this, we introduce the sparsity loss term as described in Eq. 4, where:

$$\hat{h}^v_k = \frac{1}{N}\sum_{j=1}^{N}\sigma\big(W^v_k x^v_j + b^v_k\big), \tag{29}$$

enforces an average activation $\hat{h}^v_k \approx \rho \ll 1$, directly implementing the ℓ_0-like constraint in a differentiable form. Even when input data exhibit substantial missing entries, as long as the underlying signal admits a sparse representation, a sparse autoencoder can recover that representation through its activation patterns and thereby extract the essential features.

Conclusion Sparse inputs, represented by $\|x^v_j\|_0$ in Eq. 21, naturally reside in low-dimensional semantic subspaces. By constraining the reconstruction to the nonzero support in Eqs. 23–24 and utilizing compressed sensing principles along with spark-based recoverability, as outlined in Eqs. 25–26, we ensure the existence of a unique and stable sparse code α^{v*}_j, as shown in Eq. 27. The sparse autoencoder achieves this by integrating the KL-divergence sparsity penalty with the reconstruction objective, effectively activating only the minimal set of neurons corresponding to the active dimensions. This process leads to a compact, minimally redundant representation that accurately captures the data's intrinsic subspace structure. Collectively, these theoretical foundations rigorously validate the empirical success of the Sparse Autoencoder with Adaptive Constraints (SAA) and emphasize its role in facilitating sparsity-aware multi-view representation learning.

F Limitations and Future Work

While effective in addressing cross-view sparsity variations, the current approach has several limitations that suggest directions for further improvement. To begin with, SparseMVC demonstrates strong performance across a variety of multiview data, but its effectiveness in real-world applications with noisy or incomplete scenarios remains to be fully explored. Additionally, the use of contrastive learning inherently introduces computational overhead, making it unlikely to rank among the fastest available approaches. Moreover, the framework assumes well-aligned views, whereas slight inter-view misalignment may occur in real-world scenarios.

A potential direction for future work is to incorporate structural information across views, which can be modeled through priors such as graph connectivity, inter-view relational graphs, or mutual information constraints—techniques that have been extensively explored in prior multi-view representation learning research. In contrast, the present study primarily focuses on feature-level sparsity and its adaptive regularization, rather than modeling explicit structural dependencies. Incorporating structural priors may further enhance representation quality, particularly in scenarios where inter-view relationships are semantically meaningful.