RETHINKING MULTI-VIEW CLUSTERING AS SYMMETRIC INCREMENTAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep multi-view clustering seeks shared structure from heterogeneous observations without supervision; however, existing methods often rely on shallow view-level coupling, limiting progressive cross-view reasoning. We reformulate multi-view fusion as view-wise incremental reasoning and propose VIR-SCI, which infers complementary information for each view via anchor-excluded symmetric cross-view interactions. To improve discriminability and robustness, we further introduce feature-wise incremental reasoning to adaptively modulate cross-view information. Mutual-information-constrained semantic alignment and contrastive learning jointly preserve global–local semantic coherence throughout reasoning. Extensive experiments show that VIR-SCI consistently outperforms state-of-the-art methods.

1 INTRODUCTION

Existing Deep Multi-view Clustering (DMVC) methods largely adopt static fusion schemes with limited inter-view interaction. We argue that their fundamental limitation lies not in model complexity, but in the lack of a principled cross-view reasoning mechanism, leading to shallow semantic integration. Motivated by this perspective, we rethink multi-view fusion as a progressive reasoning process.

Current DMVC research focuses on three pillars. First, SCDFL (Zhang & Che, 2024) decouples consistency and diversity, while DDMVC (Xu et al., 2025) and LUCE-CMC (Zou et al., 2024) utilize fusion-free constraints or shared encoders to enhance discriminative representations. Second, global structural awareness advances through MAGA's (Bian et al., 2024) self-attention, CSOT's (Zhang et al., 2024) optimal transport, and PNCL-FDMC's (Liu et al., 2024) neighbor-masked local structure preservation. Third, robustness is enhanced by CANDY (Guo et al., 2024) and RMCNC (Sun et al., 2024) via correspondence denoising, alongside VITAL (He et al., 2024) which employs variational category realignment.

Despite recent advances in deep multi-view clustering, existing methods still suffer from two fundamental limitations: (1) One-shot fusion with uniform view-level interactions limits effective cross-view reasoning. Most methods aggregate multi-view information in a single fusion step and model cross-view relevance mainly at the view level: Although some recent approaches introduce more granular designs, such as sample-wise consistency fusion in LFCGL (Wang et al., 2026a) and sample-adaptive view weighting in ILASC (Yang & Tan, 2026), these methods still largely rely on one-shot interactions. Similarly, methods like DWCL (Yuan et al., 2026) attempt to filter unreliable views but lack deep feature-wise reasoning mechanisms; (2) Lack of explicit mechanisms to maintain semantic coherence across representations. While representation alignment is widely adopted, existing methods often fail to enforce global–local semantic consistency after cross-view interaction or fusion. Recent works explore multi-level alignment through contrastive learning (e.g., MCMC (Bian et al., 2025), 3MC (Chen et al., 2025)) or label-distribution alignment (e.g., SSLNMVC (Yan et al., 2025)), yet they mainly focus on enforcing alignment rather than preserving semantic coherence throughout the representation transformation process.

In this paper, we propose VIR-SCI, a view-wise incremental reasoning framework for DMVC. Instead of performing one-shot fusion, our approach progressively infers complementary information for each view through anchor-excluded symmetric cross-view interactions, allowing different views to exploit heterogeneous information in an adaptive and permutation-invariant manner. Furthermore, we introduce a feature-wise incremental reasoning mechanism to selectively inject cross-view information

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

at the feature level, effectively reducing the influence of redundant or misleading information. To ensure semantic stability throughout the reasoning process, we establish a mutual-information-constrained semantic alignment as a global semantic anchor and impose a contrastive learning to maintain coherence between global and view-specific representations. Together, these components form a unified and theoretically grounded framework for robust deep multi-view clustering. In summary, our contributions can be summarized as follows:

- We reformulate multi-view clustering as a view-wise incremental reasoning problem. By introducing anchor-excluded symmetric cross-view interactions, the framework adaptively infers complementary information from other views in a permutation-invariant manner.
- We propose a hierarchical reasoning mechanism that combines semantic alignment and feature-wise incremental reasoning. This design enables selective and coherent integration of cross-view information at both view and feature levels.
- Extensive experiments on multiple benchmark datasets demonstrate that our method outperforms state-of-the-art methods. Comprehensive ablation studies and robustness analyses further validate the effectiveness and stability of each component.

2 Related Work

Based on the integration mechanism, current research of MVC can be categorized into three major categories:

Latent Space and Subspace-based Fusion. This category focuses on projecting multiple views into a shared latent space to learn a unified representation. DMCAN (Wang et al., 2026b) integrates auto-encoders with Non-negative Matrix Factorization (NMF) to learn a shared consensus matrix through element-level updates. SCMC (Fu et al., 2024) maps heterogeneous features into a joint semantic space, utilizing self-expression layers to learn consistent subspace representations. To enhance the expressive power of high-level features, SSLNMVC (Yan et al., 2025) introduces a U-Projection module that leverages feature resampling and concatenation. DWCL (Yuan et al., 2026) further refines this by employing a Best-Other (B-O) contrastive mechanism with dual-weighting to mitigate the impact of unreliable views.

Graph-based and Structural Consistency Fusion. Graph-based methods emphasize mining the topological relationships between samples for fine-grained fusion. LFCGL (Wang et al., 2026a) moves beyond coarse view-level weighting by proposing a sample-wise consistency fusion strategy, learning individual edge significance to generate a consensus graph. NSEAN (Zhao et al., 2026) jointly optimizes adaptive neighbor graph learning and nonnegative spectral embedding to automatically assign view weights based on reconstruction fidelity. To capture higher-order correlations, SMVC (Chen et al., 2024) utilizes high-order adjacency matrices within a contrastive framework to reduce feature redundancy. Additionally, MAGA (Bian et al., 2024) aggregates view-specific graph embeddings into a unified feature space via a cross-view self-attention mechanism.

Multilevel Contrastive Alignment and Robust Fusion. Recent advancements leverage self-supervised learning to achieve "soft fusion" through hierarchical semantic alignment. 3MC (Chen et al., 2025) and MCMC (Bian et al., 2025) implement comprehensive frameworks that perform contrastive learning across multiple granularities, including instance, cluster, prototype, and encoder-layer levels, to capture multi-scale representational structures. For robust fusion under noise or misalignment, STCMC-UR (Hu et al., 2025) incorporates Dempster-Shafer evidence theory to fuse belief quality and uncertainty across views. RMCNC (Sun et al., 2024) addresses noisy correspondence by maximizing a united probability to alleviate alignment bias caused by individual false positives. In scenarios with unpaired data, GLANCE (Xin et al., 2026) constructs a synthesized view as a bridge to guide global and local distribution alignment via KL divergence.

3 Method

Given a dataset with m views and n samples, it is denoted as $\{\mathbf{X}^v = \{\mathbf{x}_1^v, \ldots, \mathbf{x}_n^v\} \in \mathbb{R}^{n \times D_v}\}_{v=1}^m$, where $\mathbf{x}_i^v \in \mathbb{R}^{D_v}$ represents the i-th sample from the v-th view, and D_v denotes the original feature dimensionality of view v. Figure 1 illustrates the framework of our proposed method.

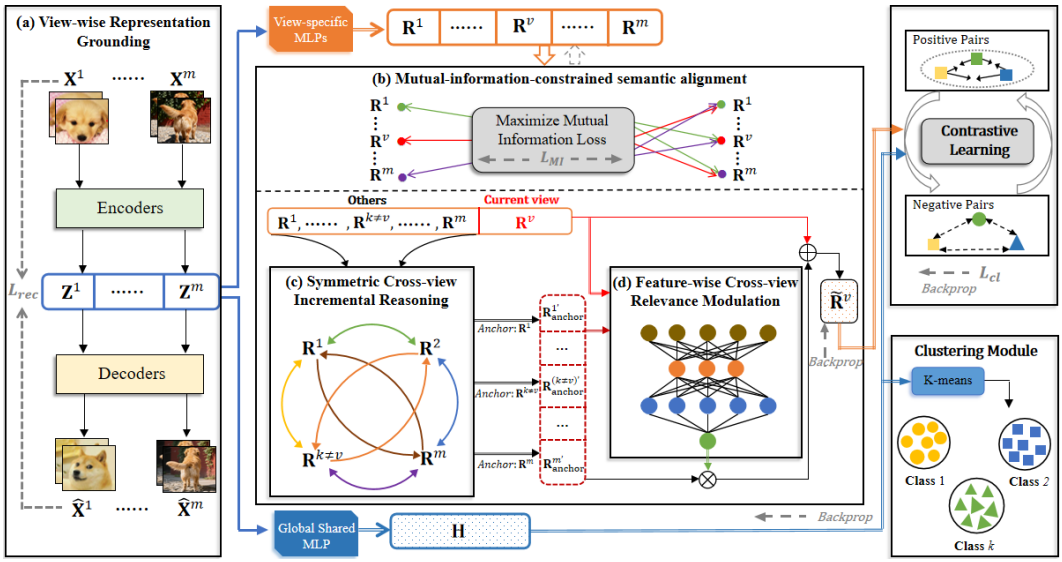


Figure 1: Overview of the proposed VIR-SCI framework. (a) **View-wise representation grounding**: view-specific autoencoders encode $\{\mathbf{X}^v\}_{v=1}^m$ into latent representations $\{\mathbf{Z}^v\}_{v=1}^m$ through reconstruction learning. (b) **Mutual-information-constrained semantic alignment**: shared and view-specific MLPs learn a global semantic representation $\mathbf{H}$ and view-specific representations $\{\mathbf{R}^v\}_{v=1}^m$, while mutual information maximization enforces cross-view semantic consistency. (c) **Symmetric cross-view incremental reasoning**: for each current view $\mathbf{R}^v$, the remaining views alternately serve as anchors and interact with the other views to infer complementary incremental information. (d) **Feature-wise cross-view relevance modulation**: a feature-wise importance map $\mathbf{F}^v$ filters and injects informative incremental components into $\mathbf{R}^v$, producing fused representations $\{\tilde{\mathbf{R}}^v\}_{v=1}^m$. The fused representations are further optimized by contrastive learning and used for clustering.

3.1 View-wise Representation Grounding

Multi-view data often contains noise and redundancy, making preprocessing important for subsequent representation learning. A common strategy is to use an encoder–decoder architecture (Hinton & Salakhutdinov, 2006) to learn compact latent representations while suppressing view-specific noise and irrelevant variations. Specifically, for the v-th view, the encoder $f_\theta^{(v)}(\cdot)$ maps the input sample $\mathbf{x}_i^v$ to the latent representation $\mathbf{z}_i^v$, and the decoder $g_\phi^{(v)}(\cdot)$ reconstructs it as $\hat{\mathbf{x}}_i^v$:

$$\mathbf{z}_i^v = f_\theta^{(v)}(\mathbf{x}_i^v), \quad \hat{\mathbf{x}}_i^v = g_\phi^{(v)}(\mathbf{z}_i^v), \tag{1}$$

where $\mathbf{z}_i^v \in \mathbb{R}^{d_v}$ and d_v denotes the dimensionality of the latent feature space for view v. Then, the reconstruction loss is defined as:

$$\mathcal{L}_{\text{rec}} = \sum_{v=1}^{m} \sum_{i=1}^{n} \|\mathbf{x}_i^v - \hat{\mathbf{x}}_i^v\|_F^2 . \tag{2}$$

3.2 Mutual-information-constrained semantic alignment

Effective cross-view reasoning requires a shared semantic space. However, the latent representations $\{\mathbf{Z}^v\}_{v=1}^m$ often contain view-specific variations that obscure clustering semantics and hinder direct cross-view reasoning. We therefore project them into a high-level semantic space to disentangle shared semantics from private factors, and impose mutual-information-constrained alignment to ensure cross-view consistency for subsequent view-wise and feature-wise reasoning. Specifically, each view-specific latent representation $\mathbf{Z}^v$ is first projected into two complementary semantic spaces to learn a global semantic representation $\mathbf{H}$ and high-level view-specific feature representations $\{\mathbf{R}^v\}_{v=1}^m$:

$$\mathbf{H} = \mathcal{F}_{\text{MLP}}\left(\left[\mathbf{Z}^1 \oplus \mathbf{Z}^2 \oplus \cdots \oplus \mathbf{Z}^m\right]; \Phi\right) \in \mathbb{R}^{n \times d}; \mathbf{R}^v = \mathcal{F}_{\text{MLP}}(\mathbf{Z}^v; \Psi^v) \in \mathbb{R}^{n \times d}, \tag{3}$$

where d denotes the dimension of the high-level semantic space, and Φ and Ψ^v represent the parameters of the global shared Multilayer Perceptron (MLP) and the view-specific MLP for the v-th view, respectively.

Although $\mathbf{R}^v$ preserves view-specific semantics in a high-level space, cross-view semantic inconsistency may remain and accumulate during fusion. To mitigate this, inspired by Oord et al. (2018), we maximize inter-view mutual information using a contrastive objective that provides a variational lower bound: $I(\mathbf{R}^v;\mathbf{R}^k) \geq -\mathcal{L}_{\mathrm{MI}}^{(v,k)}$. Here, $I(\cdot;\cdot)$ denotes mutual information. The cross-view contrastive loss between view v and view k is defined as:

$$\mathcal{L}_{\mathrm{MI}}^{(v,k)} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp\left(\mathrm{sim}(\mathbf{R}_i^v,\mathbf{R}_i^k)/\tau_1\right)}{\sum_{j=1,j\neq i}^{n}\exp\left(\mathrm{sim}(\mathbf{R}_i^v,\mathbf{R}_j^k)/\tau_1\right)}, \tag{4}$$

where τ_1 is a temperature parameter. The similarity between samples from different views is measured using cosine similarity $\mathrm{sim}(\mathbf{R}_i^v,\mathbf{R}_j^k) = \frac{(\mathbf{R}_i^v)^\top \mathbf{R}_j^k}{\|\mathbf{R}_i^v\|_2\,\|\mathbf{R}_j^k\|_2}$. By minimizing $\mathcal{L}_{\mathrm{MI}}^{(v,k)}$, the lower bound of $I(\mathbf{R}^v;\mathbf{R}^k)$ is maximized, encouraging representations of the same sample from different views to be aligned at the semantic level. The overall mutual information loss is defined as:

$$\mathcal{L}_{\mathrm{MI}} = \frac{1}{m(m-1)}\sum_{v=1}^{m}\sum_{\substack{k=1\\k\neq v}}^{m}\mathcal{L}_{\mathrm{MI}}^{(v,k)}. \tag{5}$$

The aligned $\{\mathbf{R}^v\}_{v=1}^m$ support view-wise incremental reasoning (Section 3.3), while $\mathbf{H}$ provides the semantic reference for contrastive learning (Section 3.5), forming the VIR-SCI's semantic backbone.

3.3 View-Wise Incremental Reasoning via Symmetric Cross-View Interactions

Conventional multi-view fusion strategies typically aggregate all views into a unified representation in a single step. Such one-shot fusion implicitly assumes that cross-view information is homogeneous and equally beneficial to all views, which often leads to semantic bias accumulation and loss of view-specific characteristics. In contrast, we reformulate cross-view fusion as a view-wise incremental reasoning problem, as shown in Figure 2. Specifically, for each current view $\mathbf{R}^v$, our goal is not to directly fuse all remaining views into $\mathbf{R}^v$, but to infer the complementary information that can be incrementally injected into $\mathbf{R}^v$, based on the interactions among the other views.

To this end, we define the complementary view set for the current view $\mathbf{R}^v$ as:

$$\mathcal{O}^v = \{\mathbf{R}^k \mid k = 1,\ldots,m,\ k \neq v\}. \tag{6}$$

Rather than involving $\mathbf{R}^v$ in the interaction process, we deliberately exclude the current view from cross-view reasoning, allowing the remaining views to interact symmetrically and independently of the target view. This design prevents dominant views from directly imposing biased information onto $\mathbf{R}^v$ and enables the model to extract view-agnostic incremental semantics. Within $\mathcal{O}^v$, each feature representation $\mathbf{R}^{(k\neq v)}$ is alternately treated as an anchor, while the remaining representations serve as candidates for interaction:

$$\mathcal{C}^{(k)} = \left\{\mathbf{R}^j \mid j = 1,\ldots,m,\ j \neq k,\ j \neq v\right\}. \tag{7}$$

Based on the above formulation, the anchor-specific incremental representation can be defined as follows:

$$\mathbf{R}_{\mathrm{anchor}}^{(k\neq v)\prime} = \sum_{\mathbf{R}^j\in\mathcal{C}^{(k)}}\sigma\Big(\big\langle f_a\Big(\mathbf{R}^{(k\neq v)}\Big),\ f_c\big(\mathbf{R}^j\big)\big\rangle\Big)\mathbf{R}^j, \quad \mathbf{R}^{(k\neq v)} \in \mathcal{O}^v. \tag{8}$$

In Eq.(8), $f_c(\cdot)$ denotes the candidate view embedding function, which is instantiated as the identity mapping, i.e., $f_c(\mathbf{R}^j) = \mathbf{R}^j$. The anchor embedding function $f_a(\cdot)$ is defined as an anchor-adaptive linear transformation, $f_a(\mathbf{R}^{(k\neq v)}) = \mathbf{R}^{(k\neq v)}\mathbf{W}_a$, where $\mathbf{W}_a \in \mathbb{R}^{d\times d}$ is a learnable parameter matrix that enables anchor-conditioned semantic projection. The inner product $\langle\cdot,\cdot\rangle$ measures the relevance between the anchor view and each candidate view $\mathbf{R}^j \in \mathcal{C}^{(k)}$, yielding a relevance score $s_{(k,j)} = (\mathbf{R}^j)^\top f_a\big(\mathbf{R}^{(k\neq v)}\big)$. The normalization function $\sigma(\cdot)$ implements a softmax operation over all candidate views, producing adaptive interaction weights:

$$\alpha_{(k,j)} = \frac{\exp\left(s_{(k,j)}\right)}{\sum_{\substack{j=1\\j\neq k,v}}^{m}\exp\left(s_{(k,j)}\right)}. \tag{9}$$

Finally, the anchor-specific incremental representation is obtained by aggregating candidate views weighted by their normalized relevance. By cyclically selecting each view in $\mathcal{O}^v$ as an anchor, we obtain a set of anchor-specific incremental representations $\{\mathbf{R}^{k\prime}_{\text{anchor}}\}^{m}_{k=1,\,k\neq v}$, which jointly characterize diverse complementary semantics inferred from symmetric cross-view interactions. This symmetric interaction treats all views in $\mathcal{O}^v$ equally as anchors, making the incremental representations permutation-invariant to view ordering. Different anchor choices induce diverse interaction patterns, enabling progressive exploration of cross-view complementary information beyond one-shot aggregation. Theoretical analysis is provided in Appendix A.1.1.

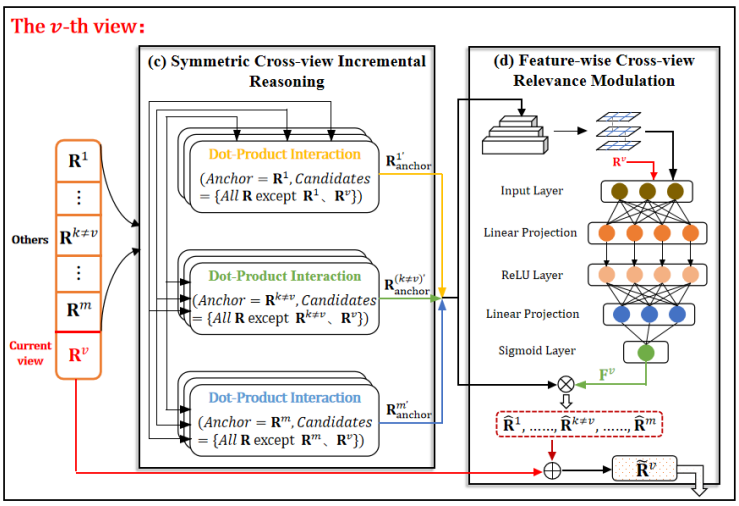


Figure 2: For the v-th view $\mathbf{R}^v$, each remaining view $\mathbf{R}^{(k\neq v)}$ is used in turn as an anchor in block (c), while the others (excluding $\mathbf{R}^v$ and $\mathbf{R}^k$) serve as candidates. Dot-product interactions estimate anchor–candidate relevance, and the weighted candidates are aggregated into an anchor-specific incremental representation $\mathbf{R}^{(k\neq v)\prime}_{\text{anchor}}$. In block (d), these representations and $\mathbf{R}^v$ generate $\mathbf{F}^v$, which adaptively filters the incremental information to obtain $\tilde{\mathbf{R}}^v$.

3.4 Feature-Wise Incremental Reasoning via Adaptive Cross-View Modulation

The symmetric cross-view interaction in Section 3.3 enables view-level incremental reasoning across complementary views. However, view-level relevance is insufficient because cross-view information may contain both informative and redundant components across feature dimensions. We therefore formulate cross-view fusion as a feature-wise incremental reasoning problem (see Figure 2), selectively injecting informative features into the current view representation.

Specifically, we first concatenate the set of anchor-specific incremental representations $\{\mathbf{R}^{k\prime}_{\text{anchor}}\}^{m}_{k=1,\,k\neq v}$ and project them into the same feature space as the current view $\mathbf{R}^v$ through a mapping MLP:

$$\mathbf{FM}_1 = \mathcal{F}_{\text{MLP}}\left(\left[\bigoplus_{k=1,\,k\neq v}^{m} \mathbf{R}^{(k\neq v)\prime}_{\text{anchor}}\right];\,\Omega\right), \tag{10}$$

where $\mathbf{FM}_1 \in \mathbb{R}^{n\times d}$ represents the cross-view incremental representation, and Ω denotes the learnable parameter of the mapping MLP. This design preserves complementary semantics while avoiding excessive feature expansion, ensuring stable optimization. To further reason about feature-level relevance, the current view $\mathbf{R}^v$ and the cross-view incremental representation $\mathbf{FM}_1$ are jointly considered to obtain a feature-wise importance map $\mathbf{F}^v$:

$$\mathbf{F}^v = \text{Sigmoid}\left(\mathcal{G}\left([\mathbf{R}^v, \mathbf{FM}_1]\right)\right), \tag{11}$$

where $\mathcal{G}(\cdot)$ denotes a multi-layer transformation with linear projections and nonlinear activations, modeling interactions between view-specific and cross-view features. The Sigmoid function constrains $\mathbf{F}^v \in \mathbb{R}^{n\times d}$ to $[0, 1]$, yielding a feature-wise importance map that assigns adaptive relevance

scores to each feature dimension. Unlike uniform or view-level weighting, this formulation captures fine-grained relevance variations, allowing the model to filter beneficial cross-view information while mitigating redundant or view-irrelevant components. Based on $\mathbf{F}^v$, selective injection of cross-view incremental information is performed as follows:

$$\widehat{\mathbf{R}}^{(k\neq v)} = \mathbf{F}^v \odot \mathbf{R}_{\text{anchor}}^{(k\neq v)\prime}, \tag{12}$$

where $\odot$ denotes element-wise multiplication. The filtered incremental representations are then fused with the current view in an additive and scale-consistent manner:

$$\tilde{\mathbf{R}}^v = \frac{1}{m}\left(\mathbf{R}^v + \sum_{\substack{k=1\\k\neq v}}^{m} \widehat{\mathbf{R}}^{(k\neq v)}\right). \tag{13}$$

This additive normalized injection preserves the semantics of the current view while selectively incorporating cross-view information, preventing auxiliary views from dominating and improving robustness to varying view quality. It also ensures bounded feature magnitudes and stable gradient propagation, as analyzed in Appendix A.1.2. The feature-wise incremental reasoning uses $\mathbf{F}^v$ to quantify feature-level contributions to cross-view fusion. Together with view-wise incremental reasoning, it forms a hierarchical framework for adaptive and discriminative multi-view representation learning.

3.5 Contrastive Learning and Clustering

View-wise and feature-wise reasoning progressively fuse complementary cross-view information, but may produce semantically inconsistent representations without explicit alignment. We therefore use contrastive learning to align the global representation $\mathbf{H}$ with the reasoned view-specific representations $\{\tilde{\mathbf{R}}^v\}_{v=1}^m$. For each sample i, $(\mathbf{H}_i, \tilde{\mathbf{R}}_i^v)$ forms a positive pair, while $(\mathbf{H}_i, \tilde{\mathbf{R}}_j^v)$ with $j \neq i$ forms a negative pair. The similarity between a pair of representations is measured using cosine similarity $\text{Sim}(\mathbf{H}_i, \tilde{\mathbf{R}}_j^v) = \frac{\langle \mathbf{H}_i, \tilde{\mathbf{R}}_j^v \rangle}{\|\mathbf{H}_i\| \cdot \|\tilde{\mathbf{R}}_j^v\|}$. The contrastive learning loss for view v is defined as:

$$\mathcal{L}_{\text{cl}}^v = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\exp\Big(\text{Sim}(\mathbf{H}_i, \tilde{\mathbf{R}}_i^v)/\tau_2\Big)}{\sum_{j=1}^{n} \exp\Big(\text{Sim}(\mathbf{H}_i, \tilde{\mathbf{R}}_j^v)/\tau_2\Big)}, \tag{14}$$

where τ_2 is a temperature parameter that controls the concentration of the similarity distribution. This loss function ensures that local reasoning processes remain consistent with the global semantic representation, thereby effectively avoiding semantic drift. Overall, the total loss of VIR-SCI is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{MI}} + \lambda_2 \mathcal{L}_{\text{cl}}, \tag{15}$$

where λ_1 and λ_2 are hyperparameters that control the relative importance of $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{cl}}$, respectively. The optimization procedure is provided in Appendix A.2. Finally, we perform k-means clustering on the global semantic representation $\mathbf{H}$:

$$Y_i = \arg\min_j \|\mathbf{H}_i - \mathbf{C}_j\|^2, \tag{16}$$

where $\mathbf{C}_j$ denotes the j-th cluster center, $Y_i \in \{1, \ldots, K\}$ is the resulting pseudo-label of the i-th sample, and K is the number of clusters.

4 Experiments

4.1 Experimental Settings

Datasets Description. Experiments are conducted on six multi-view datasets: three small-to-medium datasets, MNIST-USPS (Peng et al., 2019), NGs (Yan et al., 2025), and Caltech-5V (Fei-Fei et al., 2004); and three large-scale datasets, Fashion (Xiao et al., 2017), Cifar100 (Wu et al., 2024), and Reuters (Qin et al., 2025). Details are provided in Appendix A.3.1.

Table 1: Clustering performance of all compared methods. The best performance is highlighted in bold and the second best is underlined.

DATASETS	MNIST-USPS			NGS			CALTECH-5V			FASHION			CIFAR100			REUTERS		
METRICS	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
GCFAGG	99.56	98.71	99.56	81.20	75.85	81.20	83.36	73.31	83.36	89.28	93.14	89.28	95.97	99.35	96.05	36.20	20.30	43.60
SEM	91.72	88.27	91.72	90.00	75.51	90.00	87.14	80.25	87.14	99.29	98.19	99.29	97.30	99.44	97.98	51.70	33.95	61.36
DCMVC	89.20	86.46	89.20	62.00	56.46	62.00	82.43	72.69	82.43	96.35	93.96	96.35	98.97	98.95	98.97	45.88	30.38	56.47
HFMVC	98.10	96.00	98.00	83.93	64.73	83.93	74.79	64.49	74.79	92.68	89.76	92.68	60.74	78.21	60.29	43.24	17.70	49.50
ACCMVC	99.52	98.64	99.52	91.20	78.39	91.20	86.86	80.40	86.86	99.24	98.03	99.24	99.46	98.43	98.81	58.72	43.32	64.69
DDMVC	98.41	95.52	98.41	76.20	72.17	76.20	89.30	81.40	89.30	86.20	85.48	86.37	72.20	86.28	76.28	55.00	28.60	56.60
SSLNMVC	99.43	98.26	99.43	94.40	85.81	94.40	88.71	79.62	88.71	99.29	98.18	99.29	67.93	92.70	68.71	55.61	43.74	63.56
MGCMVC	99.70	99.10	99.70	90.00	78.91	90.00	88.70	79.50	88.70	99.30	98.20	99.30	74.31	91.15	75.28	50.27	32.81	54.30
SPARSEMVC	99.68	99.12	99.68	77.40	73.94	80.40	78.64	66.03	78.64	97.17	94.22	97.17	98.26	99.35	98.71	57.14	42.56	62.86
DMVC_MIC	96.24	89.47	96.62	82.60	75.19	82.60	84.29	72.48	84.29	89.10	93.67	89.10	95.60	98.40	96.12	65.27	43.97	68.27
NSEAN	99.24	97.86	99.24	94.20	84.95	94.20	71.79	61.87	75.86	97.50	94.89	97.50	95.46	99.21	97.03	32.86	13.44	35.59
BONE	98.20	95.20	98.20	69.00	57.90	71.60	90.60	83.10	90.60	86.81	85.46	86.82	94.59	98.78	94.59	53.35	35.62	59.07
VIR-SCI	**99.78**	**99.32**	**99.78**	**96.40**	**88.84**	**96.40**	**92.86**	**86.21**	**92.86**	**99.52**	**98.65**	**99.52**	**99.98**	**99.97**	**99.98**	**67.78**	**48.39**	**73.76**

Comparison Methods. We compare VIR-SCI with twelve recent DMVC methods. BONE (Xu et al., 2026) bridges optimization and neural networks for efficient MVC. NSEAN (Zhao et al., 2026) combines adaptive graph learning with nonnegative spectral embedding. DDMVC (Xu et al., 2025) preserves view-specific information through fusion-free consistency learning. SSLNMVC (Yan et al., 2025) aligns consensus and view-specific semantics via dual-branch supervision. MGCMVC (Ju et al., 2025) models multi-granularity representations for cross-view consistency. SparseMVC (Liu et al., 2025) handles view sparsity imbalance via adaptive sparse encoding. DMVC_MIC (Cui et al., 2025) reduces redundancy using a meta-learned information compressor. AccMVC (Yan et al., 2024) disentangles shared and view-specific semantics through anchor-based contrastive alignment. HFMVC (Jiang et al., 2024) addresses view heterogeneity in federated and decentralized settings. DCMVC (Cui et al., 2024) improves clustering discrimination with dual contrastive objectives. SEM (Xu et al., 2023) balances view reliability through adaptive reweighting. GCFAgg (Yan et al., 2023) aggregates complementary information to reduce noise and redundancy.

Evaluation Metrics. We evaluate clustering performance using clustering accuracy (ACC), normalized mutual information (NMI), and purity (PUR), which measure label consistency, shared information, and cluster homogeneity, respectively. Higher values indicate better performance.

Implementation Details. For each view, a fully connected autoencoder (Input–Fc500–Fc500–Fc2000–Fc64) extracts $\{\mathbf{Z}^v\}_{v=1}^m$. A shared MLP (Fc$(m\times 64)$–Fc256–Fc20) and a view-specific MLP (Fc64–Fc20) generate $\mathbf{H}$ and $\{\mathbf{R}^v\}_{v=1}^m$, respectively. Anchor-specific incremental representations are mapped by an MLP (Fc$((m-1)\times 64)$–Fc128–Fc20) and combined with $\mathbf{R}^v$ in a network (Input$(20+20)$–Fc500–Fc128–Fc20–Sigmoid) to produce $\mathbf{F}^v$. Experiments are conducted on an Intel i7-12700K CPU, 32GB RAM, and an NVIDIA RTX 4070 GPU (12GB), using PyTorch 2.3.1, Python 3.8.20, and CUDA 11.8. We use Adam with a learning rate of 0.0003 and batch size 256. Reconstruction is trained for 200 epochs, followed by 50 epochs of contrastive learning with dataset-specific adjustments.

4.2 COMPARATIVE RESULT ANALYSIS

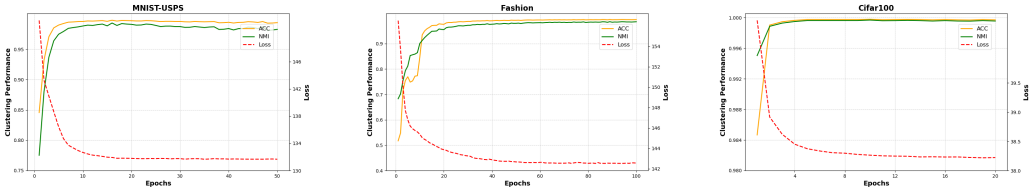
The clustering results on six datasets are reported in Table 1. Compared with BONE, which bridges optimization and neural networks without explicit consistency priors, VIR-SCI progressively integrates complementary information through symmetric cross-view interactions and feature-wise modulation, improving ACC and NMI on Caltech-5V by 2.26% and 3.11%, respectively. Unlike SSLNMVC, whose non-symmetric fusion limits interaction depth, VIR-SCI performs fully symmetric incremental reasoning with feature-wise selection, yielding a 3.03% NMI gain on NGs. Although MGCMVC and DMVC_MIC reduce redundancy through adaptive weighting and meta-learned compression, respectively, their integration of complementary semantics remains limited. VIR-SCI surpasses MGCMVC by 0.08% and 0.22% ACC on MNIST-USPS and Fashion, respectively, and outperforms DMVC_MIC by 2.51% ACC on Reuters. On Cifar100, VIR-SCI further exceeds the corresponding second-best results by 1.01% PUR. Overall, VIR-SCI achieves the best performance across all

six datasets, demonstrating the effectiveness of symmetric incremental reasoning and feature-wise selective integration.

4.3 Model Analysis

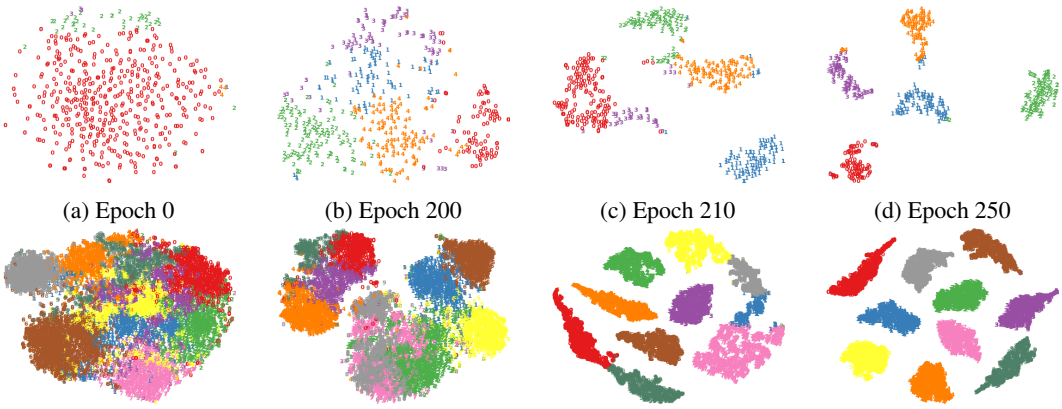
Convergence Analysis. As shown in Figure 3, the training loss decreases rapidly and converges, while ACC and NMI increase and stabilize with minor fluctuations, indicating stable optimization of VIR-SCI. Results on other datasets are provided in Appendix A.3.2.

Visualization Analysis of Feature Clustering Process. We apply t-SNE (Maaten & Hinton, 2008) to visualize $\mathbf{H}$ across training epochs. As shown in Figure 4, clusters become progressively more compact and well separated, demonstrating the discriminative capability of VIR-SCI. Results on other datasets are provided in Appendix A.3.3.



(a) MNIST-USPS (b) Fashion (c) Cifar100

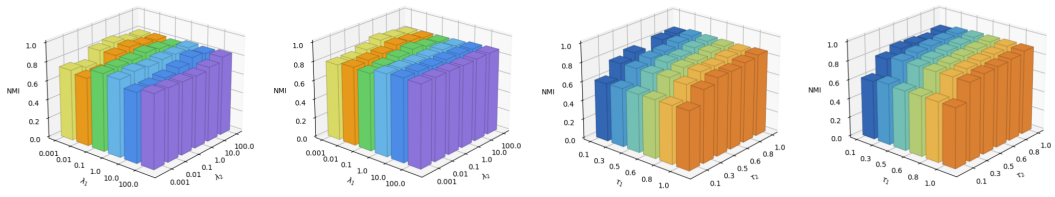
Figure 3: Clustering metrics (ACC, NMI) and loss curves on MNIST-USPS, Fashion, and Cifar100.



(a) Epoch 0 (b) Epoch 200 (c) Epoch 210 (d) Epoch 250

(e) Epoch 0 (f) Epoch 200 (g) Epoch 210 (h) Epoch 250

Figure 4: T-SNE clustering visualizations on NGs ((a)–(d)) and Fashion ((e)–(h)) across epochs.



(a) Caltech-5V (λ_1, λ_2) (b) NGs (λ_1, λ_2) (c) Caltech-5V (τ_1, τ_2) (d) NGs (τ_1, τ_2)

Figure 5: Parameter sensitivity analysis using NMI on Caltech-5V and NGs.

Parameter Sensitivity Analysis. We analyze the loss weights λ_1, λ_2 (Eq. (15)) and temperatures τ_1, τ_2 (Eqs. (4), (14)) in Figure 5; results on other datasets are provided in Appendix A.3.4. With $\tau_1 = 0.5$ and $\tau_2 = 1.0$, we vary λ_1 and λ_2 from 10^{-3} to 100 and set $\lambda_1 = 0.5$, $\lambda_2 = 1.0$. Varying

Table 2: Ablation results on Caltech-5V, NGs, Reuters and Cifar100 datasets.

Model	L_{rec}	L_{MI}	$L_{\text{cl}}^{\mathbf{R}^v}$	$L_{\text{cl}}^{+\text{S}}$	$L_{\text{cl}}^{+\text{F}}$	$L_{\text{cl}}^{+\text{A}}$	Caltech-5V			NGs			Reuters			Cifar100		
							ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
M_1	✓						68.50	56.31	71.64	55.40	28.69	55.40	42.11	17.37	45.46	96.36	98.32	97.19
M_2		✓					73.21	63.77	73.21	51.80	36.39	56.00	55.25	33.58	56.52	79.95	90.29	82.82
M_3			✓				82.14	72.18	82.14	53.60	28.74	53.60	60.07	40.49	66.05	62.54	88.95	70.95
M_4			✓	✓	✓	✓	84.36	74.89	84.36	53.60	31.79	53.60	60.35	40.55	66.32	78.41	93.88	83.40
M_5	✓	✓					77.57	63.79	77.57	92.40	79.82	92.40	62.25	38.08	65.96	99.97	99.96	99.97
M_6	✓		✓				87.86	79.06	87.86	87.80	72.58	87.80	60.83	42.56	65.39	91.95	97.87	93.80
M_7	✓		✓	✓	✓	✓	89.36	80.70	89.36	90.40	78.40	90.40	66.13	44.18	66.13	99.85	99.83	99.85
M_8		✓	✓				87.07	78.43	87.07	55.00	34.13	55.00	62.10	42.73	63.72	99.95	99.93	99.95
M_9		✓	✓	✓	✓	✓	88.14	79.18	88.14	57.80	35.73	57.80	62.11	42.90	66.76	99.98	99.96	99.98
M_{10}	✓	✓	✓	✓			80.64	67.90	80.64	94.00	82.84	94.00	62.92	41.88	64.80	99.97	99.96	99.97
M_{11}	✓		✓	✓			80.57	67.21	80.57	94.40	84.16	94.40	59.47	38.34	61.34	99.96	99.95	99.96
M_{12}	✓	✓	✓		✓		85.86	75.06	85.86	95.80	88.07	95.80	63.74	42.40	65.06	99.97	99.96	99.97
M_{13}	✓		✓		✓		85.43	74.65	85.43	94.60	84.43	94.60	58.77	37.85	61.11	99.96	99.95	99.96
M_{14}	✓	✓	✓			✓	91.43	84.07	91.43	96.00	88.21	96.00	65.71	43.78	65.72	99.97	99.96	99.97
M_{15}	✓		✓			✓	91.07	83.72	91.07	95.00	85.19	95.00	62.54	42.19	63.56	98.47	99.46	98.76
M_{16}	✓	✓	✓	✓	✓		86.86	75.18	86.86	96.00	88.21	96.00	64.73	43.48	65.16	99.98	99.96	99.98
M_{17}	✓		✓	✓	✓		86.71	74.66	86.71	94.80	85.05	94.80	58.73	37.89	61.08	99.97	99.96	99.97
M_{18}	✓	✓	✓	✓		✓	91.71	84.37	91.71	96.20	88.40	96.20	65.20	42.74	66.04	99.97	99.96	99.97
M_{19}	✓		✓	✓		✓	91.14	83.89	91.14	95.00	85.19	95.00	62.46	42.22	63.65	98.47	99.46	98.76
M_{20}	✓	✓	✓		✓	✓	91.50	83.60	91.50	96.20	88.40	96.20	65.07	42.15	65.81	99.98	99.96	99.98
M_{21}	✓		✓		✓	✓	91.43	83.37	91.43	94.80	85.17	94.80	63.73	42.34	63.79	99.80	99.73	99.80
VIR-SCI	✓	✓	✓	✓	✓	✓	**92.86**	**86.21**	**92.86**	**96.40**	**88.84**	**96.40**	**67.78**	**48.39**	**73.76**	**99.98**	**99.97**	**99.98**

τ_1 and τ_2 over $\{0.1, 0.3, 0.5, 0.6, 0.8, 1.0\}$ yields stable performance at larger values, indicating low sensitivity; we thus use $\tau_1 = 0.5$, $\tau_2 = 1.0$ by default.

Ablation Study. To evaluate VIR-SCI components, we conduct ablation studies on all datasets (Table 2), with MNIST-USPS and Fashion results in Appendix A.3.5. Here, L_{rec} is the reconstruction loss, L_{MI} the mutual information loss, and $L_{\text{cl}}^{\mathbf{R}^v}$ the contrastive loss between $\mathbf{R}^v$ and $\mathbf{H}$. $L_{\text{cl}}^{+\text{S}}$ and $L_{\text{cl}}^{+\text{F}}$ denote symmetric cross-view incremental reasoning and feature-wise relevance modulation applied to $\mathbf{R}^v$, respectively, while $L_{\text{cl}}^{+\text{A}}$ represents additive fusion of the current view with the incremental information. M_1–M_4 are single-loss models with poor clustering performance. Removing L_{rec} (M_9) causes a substantial drop (e.g., 38.6% ACC on NGs), highlighting its role in preserving data structure and stabilizing learning. Excluding L_{MI} (M_7) also reduces performance (e.g., 3.5% ACC on Caltech-5V), confirming its importance for cross-view semantic consistency. Contrastive learning alone (M_1 vs. M_6) yields notable gains, while comparisons such as M_3 vs. M_4, M_6 vs. M_7, and M_8 vs. M_9 demonstrate the overall effectiveness of incremental reasoning modules. Then, we provide a detailed analysis. Replacing symmetric cross-view incremental reasoning with simple mean aggregation (M_{20}), leading to one-shot fusion and reduced complementary information exploitation (NMI drops 6.24% on Reuters). Without feature-wise relevance modulation (M_{18}), all feature dimensions are treated equally. Irrelevant features may interfere with the current view, resulting in a 1.15% ACC drop on Caltech-5V. Excluding the current view in fusion (M_{16}) overly relies on auxiliary views, reducing robustness. Sequentially removing these modules (M_{10}, M_{14}, M_{12}) confirms that symmetric reasoning, feature-wise modulation, and additive fusion are all essential. Smaller changes on Cifar100 occur as performance is near saturation. Other variants without L_{MI} (M_{11}, M_{13}, M_{15}, M_{17}, M_{19}, M_{21}) further highlight its role. Overall, the ablation results show that components of VIR-SCI jointly contribute to improved clustering performance.

5 Conclusion

In this paper, we proposed VIR-SCI, a multi-view clustering framework that integrates mutual-information-constrained semantic alignment with view-wise and feature-wise incremental reasoning. VIR-SCI effectively infers and fuses complementary cross-view information while maintaining semantic consistency. Extensive experiments validate its effectiveness. A limitation is that our evaluation mainly focuses on datasets with moderate feature dimensionality, leaving its scalability to ultra-high-dimensional multi-view data underexplored. Future work will investigate more advanced fusion strategies and extend VIR-SCI to high-dimensional scenarios such as multi-omics data analysis.

AI USE STATEMENT

In this work, we used generative AI tools only to improve the grammar, readability, clarity, and language consistency of the manuscript. We did not use generative AI tools for any tasks requiring mandatory disclosure. All AI-assisted edits were carefully reviewed by the authors. Generative AI tools were not used to generate research ideas, formulate theoretical claims, design the methodology or experiments, construct proofs, implement the proposed method, or analyze and interpret the experimental results. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on methodological research for unsupervised multi-view clustering and does not involve human subjects, personally identifiable information, or interventions with individuals. Experiments are conducted on standard benchmark datasets following their established research usage. We are not aware of any significant ethical risks arising directly from the proposed method. As with other general-purpose representation learning techniques, potential downstream applications should follow appropriate data privacy, fairness, and domain-specific ethical requirements.

REPRODUCIBILITY STATEMENT

We provide the model architecture, optimization settings, hyperparameters, and training details in the main paper. Detailed dataset descriptions, additional experimental results, the optimization procedure (Appendix A.2), and theoretical analyses are provided in the appendix. Ablation and parameter sensitivity studies further document the contributions and robustness of individual components. We also provide the source code as supplementary material to facilitate reproduction of the reported results.

REFERENCES

Jintang Bian, Xiaohua Xie, Jian-Huang Lai, and Feiping Nie. Multi-view contrastive clustering via integrating graph aggregation and confidence enhancement. *Information Fusion*, 108:102393, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2024.102393.

Jintang Bian, Yixiang Lin, Xiaohua Xie, Chang-Dong Wang, Lingxiao Yang, Jian-Huang Lai, and Feiping Nie. Multilevel contrastive multiview clustering with dual self-supervised learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):10422–10436, 2025. doi: 10.1109/TNNLS.2025.3552969.

Bowei Chen, Sen Xu, Heyang Xu, Xuesheng Bian, Naixuan Guo, Xiufang Xu, Xiaopeng Hua, and Tian Zhou. Structural deep multi-view clustering with integrated abstraction and detail. *Neural Networks*, 175:106287, 2024.

Zhe Chen, Xiao-Jun Wu, Tianyang Xu, Hui Li, and Josef Kittler. Multi-layer multi-level comprehensive learning for deep multi-view clustering. *Information Fusion*, 116:102785, 2025.

Jinrong Cui, Yuting Li, Han Huang, and Jie Wen. Dual contrast-driven deep multi-view clustering. *IEEE Transactions on Image Processing*, 33:4753–4764, 2024. doi: 10.1109/TIP.2024.3444269.

Jinrong Cui, Bang Liufu, Chongjie Dong, Jingcheng Ke, and Jie Wen. Deep multi-view clustering with meta information compression. *IEEE Transactions on Image Processing*, 34:7514–7527, 2025. doi: 10.1109/TIP.2025.3630899.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Lele Fu, Sheng Huang, Lei Zhang, Jinghua Yang, Zibin Zheng, Chuanfu Zhang, and Chuan Chen. Subspace-contrastive multi-view clustering. *ACM Transactions on Knowledge Discovery from Data*, 18(9):1–35, 2024.

Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. Robust contrastive multi-view clustering against dual noisy correspondence. *Advances in Neural Information Processing Systems*, 37:121401–121421, 2024.

Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. Robust variational contrastive learning for partially view-unaligned clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4167–4176, 2024.

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

Shizhe Hu, Binyan Tian, Weibo Liu, and Yangdong Ye. Self-supervised trusted contrastive multi-view clustering with uncertainty refined. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17305–17313, 2025.

Xiaorui Jiang, Zhongyi Ma, Yulin Fu, Yong Liao, and Pengyuan Zhou. Heterogeneity-aware federated deep multi-view clustering towards diverse feature representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9184–9193, 2024.

Hengrong Ju, Yang Lu, Weiping Ding, Wei Zhang, and Xibei Yang. Multigranularity information fused contrastive learning with multiview clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):18915–18929, 2025. doi: 10.1109/TNNLS.2025.3574885.

Mingyang Liu, Zuyuan Yang, Wei Han, and Shengli Xie. Progressive neighbor-masked contrastive learning for fusion-style deep multi-view clustering. *Neural Networks*, 179:106503, 2024.

Ruimeng Liu, Xin Zou, Chang Tang, Xiao Zheng, Xingchen Hu, Kun Sun, and Xinwang Liu. Sparsemvc: Probing cross-view sparsity variations for multi-view clustering. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 154594–154622. Curran Associates, Inc., 2025. doi: 10.52202/085713-5172.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *International conference on machine learning*, pp. 5092–5101. PMLR, 2019.

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

Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):9150–9162, 2024. doi: 10.1109/TKDE.2024.3423307.

Bocheng Wang, Jiawei Wang, Chusheng Zeng, and Mulin Chen. Locality-driven flexible consensus graph learning for multi-view clustering. *Pattern Recognition*, 169:111853, 2026a.

Dexian Wang, Sha Yang, Ping Deng, Zheng Yu, Qiaofeng Wu, Wei Huang, Pengfei Zhang, and Weiping Ding. Deep multi-view clustering algorithm with integrated auto-encoder and non-negative matrix factorization. *Applied Soft Computing*, pp. 114552, 2026b.

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*, 26:9150–9162, 2024. doi: 10.1109/TMM.2024.3387298.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Like Xin, Wanqi Yang, Lei Wang, and Ming Yang. Global-and-local guidance with synthesized view for unpaired multi-view clustering. *Information Processing & Management*, 63(3):104502, 2026.

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

Jie Xu, Shuo Chen, Yazhou Ren, Xiaoshuang Shi, Hengtao Shen, Gang Niu, and Xiaofeng Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. *Advances in neural information processing systems*, 36:1119–1131, 2023.

Junpeng Xu, Min Meng, Jigang Liu, and Jigang Wu. Deep multi-view clustering with diverse and discriminative feature learning. *Pattern Recognition*, 161:111322, 2025.

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

Weiqing Yan, Yuanyang Zhang, Chang Tang, Wujie Zhou, and Weisi Lin. Anchor-sharing and cluster-wise contrastive network for multiview representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3797–3807, 2024.

Weiqing Yan, Tingyu Yang, and Chang Tang. Self-supervised semantic soft label learning network for deep multi-view clustering. *IEEE Transactions on Multimedia*, 27:4971–4983, 2025.

Zengbiao Yang and Yihua Tan. Deep multi-view clustering based on instance-level adaptive structural contrastive learning. *Neural Networks*, pp. 108581, 2026.

Hanning Yuan, Zhihui Zhang, Qi Guo, Lianhua Chi, Sijie Ruan, Wei Zhou, Jinhui Pang, and Xiaoshuai Hao. Dwcl: Dual-weighted contrastive learning for robust multi-view clustering. *Engineering Applications of Artificial Intelligence*, 165:113532, 2026.

Fenghua Zhang and Hangjun Che. Separable consistency and diversity feature learning for multi-view clustering. *IEEE Signal Processing Letters*, 31:1595–1599, 2024.

Qian Zhang, Lin Zhang, Ran Song, Runmin Cong, Yonghuai Liu, and Wei Zhang. Learning common semantics via optimal transport for contrastive multi-view clustering. *IEEE Transactions on Image Processing*, 33:4501–4515, 2024. doi: 10.1109/TIP.2024.3436615.

Mingyu Zhao, Feiping Nie, Cong Wang, Xuelong Li, Zehan Tan, and Huaqiang Hu. Nonnegative spectral embedding learning with adaptive neighbors for multi-view clustering. *Neural Networks*, pp. 108537, 2026.

Guoliang Zou, Yangdong Ye, Tongji Chen, and Shizhe Hu. Learning dual enhanced representation for contrastive multi-view clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8731–8739, 2024.

A Appendix

A.1 Theoretical Analysis

A.1.1 Theoretical Analysis of Progressive Information Coverage

To theoretically analyze the effectiveness of the proposed symmetric cross-view incremental reasoning mechanism, we investigate its capability to progressively expand the coverage of cross-view complementary information. Unlike conventional one-shot fusion strategies that directly aggregate all complementary views into a single representation, our method infers incremental semantics by inducing diverse interaction patterns conditioned on different anchor views. This analysis aims to show that such anchor-conditioned reasoning yields a richer and more comprehensive exploration of the cross-view incremental information space. For a given current view representation $\mathbf{R}^v$, let $\mathcal{O}^v = \{\mathbf{R}^k \mid k = 1, \ldots, m,\ k \neq v\}$ denote the set of complementary view representations.

Theorem 1. *Assume that all view representations lie in a bounded vector space and that selecting different complementary views as anchors induces non-identical incremental reasoning outcomes. Then, by cyclically selecting each element in $\mathcal{O}^v$ as an anchor, the proposed symmetric cross-view incremental reasoning mechanism produces a set of anchor-specific incremental representations $\{\mathbf{R}^{k\prime}_{\text{anchor}}\}^m_{k=1,\,k\neq v}$, whose spanned subspace strictly covers a larger portion of the cross-view incremental information space than that obtained by any single aggregated fusion over $\mathcal{O}^v$.*

Proof. We first consider a conventional one-shot fusion strategy that aggregates all complementary views into a single representation:

$$\bar{\mathbf{R}} = \sum_{\substack{k=1 \\ k\neq v}}^{m} \beta_k \mathbf{R}^k, \tag{17}$$

where $\{\beta_k\}$ are fixed or globally normalized coefficients. By construction, the resulting representation $\bar{\mathbf{R}}$ lies in a one-dimensional linear subspace, i.e., $\text{span}(\bar{\mathbf{R}}) = \left\{\lambda\bar{\mathbf{R}} \mid \lambda \in \mathbb{R}\right\}$. Therefore, regardless of how the coefficients $\{\beta_k\}^m_{k=1,k\neq v}$ are chosen, the dimensionality of the information captured by such a fusion strategy satisfies:

$$\dim\big(\text{span}(\bar{\mathbf{R}})\big) = 1. \tag{18}$$

However, our proposed symmetric cross-view incremental reasoning mechanism does not directly fuse $\mathcal{O}^v$ into a single representation. Instead, each complementary view representation $\mathbf{R}^{(k\neq v)} \in \mathcal{O}^v$ is alternately selected as an anchor, while the remaining $(m-2)$ representations act as candidates for interaction. Conditioned on the anchor, an anchor-adaptive transformation $f_a(\mathbf{R}^{(k\neq v)}) = \mathbf{R}^{(k\neq v)}\mathbf{W}_a$ is first applied, and interaction weights are computed as:

$$\alpha_{(k,j)} = \frac{\exp\left(\left(\mathbf{R}^j\right)^\top f_a(\mathbf{R}^{(k\neq v)})\right)}{\sum_{\substack{j=1 \\ j\neq k,v}}^{m} \exp\left(\left(\mathbf{R}^j\right)^\top f_a(\mathbf{R}^{(k\neq v)})\right)}. \tag{19}$$

The corresponding anchor-specific incremental representation is then defined as: $\mathbf{R}^{(k\neq v)\prime}_{\text{anchor}} = \sum_{\mathbf{R}^j \in \mathcal{C}^{(k,j)}} \alpha_{(k,j)}\mathbf{R}^j$.

Since the interaction weights explicitly depend on the selected anchor, different anchors induce distinct reasoning trajectories over the same candidate set. Under the non-degeneracy assumption, there exist at least two anchors $\mathbf{R}^{k_1} \neq \mathbf{R}^{k_2}$ such that $\mathbf{R}^{(k_1\neq v)\prime}_{\text{anchor}} \neq \mathbf{R}^{(k_2\neq v)\prime}_{\text{anchor}}$. As a result, the corresponding incremental representations are linearly independent, yielding:

$$\dim\Big(\text{span}\left(\{\mathbf{R}^{(k\neq v)\prime}_{\text{anchor}}\}^m_{k=1,k\neq v}\right)\Big) \geq 2. \tag{20}$$

Combining equation 18 and equation 20, we obtain:

$$\dim\big(\text{span}\left(\{\mathbf{R}^{k\prime}_{\text{fused}}\}^m_{k=1,k\neq v}\right)\big) > \dim\big(\text{span}(\bar{\mathbf{R}})\big), \tag{21}$$

which means anchor-conditioned incremental reasoning covers a strictly larger subspace of cross-view incremental information than any single aggregated fusion. By cycling through all complementary views as anchors, the proposed mechanism progressively expands the explored semantic space, thereby enabling more comprehensive cross-view information coverage. □

A.1.2 Theoretical Analysis of Gradient Stability

To theoretically justify that the proposed feature-wise incremental reasoning mechanism ensures stable gradient propagation during multi-view fusion, we conduct the following analysis. Specifically, the locally reasoned view-specific representation for the current view $\mathbf{R}^v$ is defined as:

$$\tilde{\mathbf{R}}^v = \frac{1}{m}\left(\mathbf{R}^v + \sum_{\substack{k=1 \\ k\neq v}}^{m} \mathbf{F}^v \odot \mathbf{R}^{(k\neq v)\prime}_{\text{anchor}}\right), \tag{22}$$

where the feature-wise importance map $\mathbf{F}^v = \text{Sigmoid}\left(\text{Linear}_2, \text{ReLU}\left(\text{Linear}_1[\mathbf{R}^v, \mathbf{FM}_1]\right)\right)$.

Theorem 2. *Assume that all view representations $\{\mathbf{R}^v\}_{v=1}^m$ are bounded and non-trivial, and that the parameters of all linear layers and MLPs in this mechanism have finite Frobenius norms. Then, for any differentiable loss function $\mathcal{L}$, the gradient with respect to the current view representation satisfies:*

$$\lambda_{\min}(\mathbf{J}_{\mathrm{FM}}) \left\| \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{R}}^v} \right\|_F \leq \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{R}^v} \right\|_F \leq \|\mathbf{J}_{\mathrm{FM}}\|_F \left\| \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{R}}^v} \right\|_F, \tag{23}$$

where $\mathbf{J}_{\mathrm{FM}}$ denotes the local Jacobian matrix of the feature-wise cross-view relevance modulation with respect to $\mathbf{R}^v$. Under non-saturated activation conditions, $\lambda_{\min}(\mathbf{J}_{\mathrm{FM}}) > 0$ and $\|\mathbf{J}_{\mathrm{FM}}\|_F < \infty$, ensuring non-vanishing and bounded gradient propagation.

Proof. By the chain rule, we have:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{R}^v} = \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{R}}^v} \cdot \frac{\partial \tilde{\mathbf{R}}^v}{\partial \mathbf{R}^v}. \tag{24}$$

From Eq. (22), the local Jacobian matrix of the feature-wise cross-view relevance modulation with respect to $\mathbf{R}^v$ is given by:

$$\mathbf{J}_{\mathrm{FM}} := \frac{\partial \tilde{\mathbf{R}}^v}{\partial \mathbf{R}^v} = \frac{1}{m} \left(\mathbf{I} + \sum_{k=1, k\neq v}^{m} \mathrm{diag}\left(\mathbf{R}_{\mathrm{anchor}}^{(k\neq v)\prime} \right) \cdot \frac{\partial \mathbf{F}^v}{\partial \mathbf{R}^v} \right). \tag{25}$$

The feature-wise importance map $\mathbf{F}^v$ is generated through the following transformation chain:

$$\mathbf{R}^v \xrightarrow{\mathrm{Linear}_1,\,\mathrm{ReLU}} \mathbf{FM}_2 \xrightarrow{\mathrm{Linear}_2,\,\mathrm{Sigmoid}} \mathbf{F}^v. \tag{26}$$

Applying the chain rule along this transformation path yields:

$$\frac{\partial \mathbf{F}^v}{\partial \mathbf{R}^v} = \underbrace{\frac{\partial \mathbf{F}^v}{\partial\big(\mathrm{Linear}_2(\mathbf{FM}_2)\big)}}_{\text{(a) Sigmoid Jacobian}} \cdot \underbrace{\frac{\partial\big(\mathrm{Linear}_2(\mathbf{FM}_2)\big)}{\partial \mathbf{FM}_2}}_{\text{(b) linear layer } \mathbf{W}_2} \cdot \underbrace{\frac{\partial \mathbf{FM}_2}{\partial\big(\mathrm{Linear}_1([\mathbf{R}^v, \mathbf{FM}_1])\big)}}_{\text{(c) ReLU Jacobian}} \cdot \underbrace{\frac{\partial\big(\mathrm{Linear}_1([\mathbf{R}^v, \mathbf{FM}_1])\big)}{\partial \mathbf{R}^v}}_{\text{(d) submatrix } \mathbf{W}_1^{(v)}}, \tag{27}$$

where $\mathbf{W}_2$ is the linear transformation; $\mathbf{W}_1^{(v)}$ is the submatrix corresponding to the current view $\mathbf{R}^v$.

Under non-saturated activation regions, the Jacobians of ReLU and Sigmoid are bounded, and the fused representations $\{\mathbf{R}_{\mathrm{anchor}}^{k\prime}\}_{k=1,k\neq v}^m$ are bounded due to convex combination with finite coefficients $\alpha_{(k,j)}$. Consequently, $\mathbf{J}_{\mathrm{FM}}$ is positive definite with a strictly positive minimum eigenvalue and a finite Frobenius norm:

$$0 < \lambda_{\min}(\mathbf{J}_{\mathrm{FM}}) \leq \|\mathbf{J}_{\mathrm{FM}}\|_F < \infty. \tag{28}$$

Therefore, the gradient bounds in Eq. (23) hold, guaranteeing that gradient propagation through the feature-wise cross-view relevance modulation is locally stable. □

A.2 Optimization Procedure

The optimization procedure is summarized in Algorithm 1.

A.3 Additional Experimental Results

A.3.1 Datasets Description

Our experiments are conducted on six widely used multi-view datasets, including MNIST-USPS (Peng et al., 2019), NGs (Yan et al., 2025), Caltech-5V (Fei-Fei et al., 2004), Fashion (Xiao et al., 2017), Cifar100 (Wu et al., 2024), and Reuters (Qin et al., 2025). Among them, MNIST-USPS, NGs, and Caltech-5V are small-sample datasets, while Fashion, Cifar100, and Reuters are large-scale datasets. Detailed information of these datasets is summarized in Table 3.

Algorithm 1 Optimization of VIR-SCI

Input: Multi-view data $\{\mathbf{X}^v\}_{v=1}^m$; τ_1, τ_2; λ_1, λ_2.
Output: Cluster labels $\mathbf{Y}$.

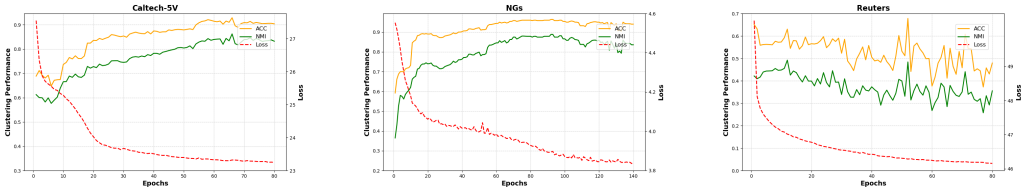
Pretrain $f_\theta^{(v)}(\cdot)$ and $g_\phi^{(v)}(\cdot)$ by minimizing Eq. (2).
while not reach the maximum iteration $T_{\max}$ **do**
  Compute $\mathbf{H}$ and $\{\mathbf{R}^v\}_{v=1}^m$ according to Eq. (3).
  Compute $\mathcal{L}_{\mathrm{MI}}$ according to Eq. (5).
  for $v = 1$ **to** m **do**
    Compute $\{\mathbf{R}_{\mathrm{anchor}}^{k\prime}\}_{k=1,\,k\neq v}^m$ according to Eq. (8).
    Compute $\mathbf{F}^v$ according to Eq. (11).
    Compute $\widehat{\mathbf{R}}^{(k\neq v)}$ according to Eq. (12).
    Obtain $\tilde{\mathbf{R}}^v$ according to Eq. (13).
  end for
  Compute $\mathcal{L}_{\mathrm{cl}}$ according to Eq. (14).
  Update all network parameters via backpropagation using the Adam optimizer.
end while
Apply k-means clustering on $\mathbf{H}$ to obtain the final cluster assignments according to Eq. (16).

Table 3: Information of the datasets used in our experiments.

Dataset	Samples	Views	Classes	View Dimensions
MNIST-USPS	5000	2	10	784/784
NGs	500	3	5	2000/2000/2000
Caltech-5V	1400	5	7	40/254/928/512/1984
Fashion	10000	3	10	784/784/784
CIFAR100	50000	3	100	512/2048/1024
Reuters	18758	5	6	10/10/10/10/10

A.3.2 Additional Convergence Analysis

Figure 6 presents the convergence behavior of VIR-SCI on the Caltech-5V, NGs, and Reuters datasets. Consistent with the main text, the training loss rapidly converges on all datasets. On Caltech-5V and NGs, ACC and NMI gradually stabilize in the later training stages, whereas some fluctuations are observed on Reuters. This phenomenon can be attributed to the inherent characteristics of Reuters, including severe class imbalance and substantial semantic overlap among categories, which make clustering metrics highly sensitive to representation updates near decision boundaries. Nevertheless, the overall convergence trend remains stable, indicating reliable optimization behavior of VIR-SCI across different datasets.



(a) Caltech-5V (b) NGs (c) Reuters

Figure 6: Clustering metrics (ACC, NMI) and loss curves over training epochs on the Caltech-5V, NGs, and Reuters datasets.

Table 4: Ablation results on MNIST-USPS and Fashion datasets.

Model	L_{rec}	L_{MI}	$L_{\text{cl}}^{\mathbf{R}^v}$	$L_{\text{cl}}^{+\text{S}}$	$L_{\text{cl}}^{+\text{F}}$	$L_{\text{cl}}^{+\text{A}}$	MNIST-USPS			Fashion		
							ACC	NMI	PUR	ACC	NMI	PUR
M_1	✓						54.72	43.67	54.72	67.60	66.03	72.21
M_2		✓					98.86	97.16	98.86	86.82	86.75	86.82
M_3			✓				99.46	98.45	99.46	98.84	97.25	98.84
M_4			✓	✓	✓	✓	99.66	98.96	99.66	99.17	97.95	99.17
M_5	✓	✓					99.46	98.41	99.46	99.24	97.99	99.24
M_6	✓		✓				99.58	98.74	99.58	99.17	97.92	99.17
M_7	✓		✓	✓	✓	✓	99.66	98.99	99.66	99.45	98.48	99.45
M_8		✓	✓				99.52	98.54	99.52	99.27	98.09	99.27
M_9		✓	✓	✓	✓	✓	99.64	98.90	99.64	99.37	98.36	99.37
M_{10}	✓	✓	✓	✓			99.68	99.03	99.68	99.04	97.55	99.04
M_{11}	✓		✓	✓			99.64	98.91	99.64	99.07	97.56	99.07
M_{12}	✓	✓	✓		✓		99.76	99.32	99.76	99.02	97.56	99.02
M_{13}	✓		✓		✓		99.64	98.92	99.64	98.92	97.36	98.92
M_{14}	✓	✓	✓			✓	99.66	98.97	99.66	99.45	98.51	99.45
M_{15}	✓		✓			✓	99.62	98.89	99.62	99.42	98.43	99.42
M_{16}	✓	✓	✓	✓	✓		99.78	99.32	99.78	99.03	97.52	99.03
M_{17}	✓		✓	✓	✓		99.64	98.92	99.64	99.02	97.49	99.02
M_{18}	✓	✓	✓	✓		✓	99.66	98.97	99.66	99.45	98.51	99.45
M_{19}	✓		✓	✓		✓	99.62	98.89	99.62	99.42	98.43	99.42
M_{20}	✓	✓	✓		✓	✓	99.74	99.20	99.74	99.43	98.49	99.43
M_{21}	✓		✓		✓	✓	99.60	98.81	99.60	99.44	98.49	99.44
VIR-SCI	✓	✓	✓	✓	✓	✓	**99.78**	**99.32**	**99.78**	**99.52**	**98.65**	**99.52**

A.3.3 Additional Visualization Analysis of Feature Clustering Process

We provide additional t-SNE visualizations of the global semantic representation $\mathbf{H}$ at various training epochs for other datasets in Figure 7. The results are consistent with the observations in the main text: as training progresses, distinct clusters gradually form and evolve into compact, well-separated structures. These additional visualizations further confirm that the proposed fusion modules and training paradigm consistently enhance feature discrimination and clustering performance across different datasets.

A.3.4 Additional Parameter Sensitivity Analysis

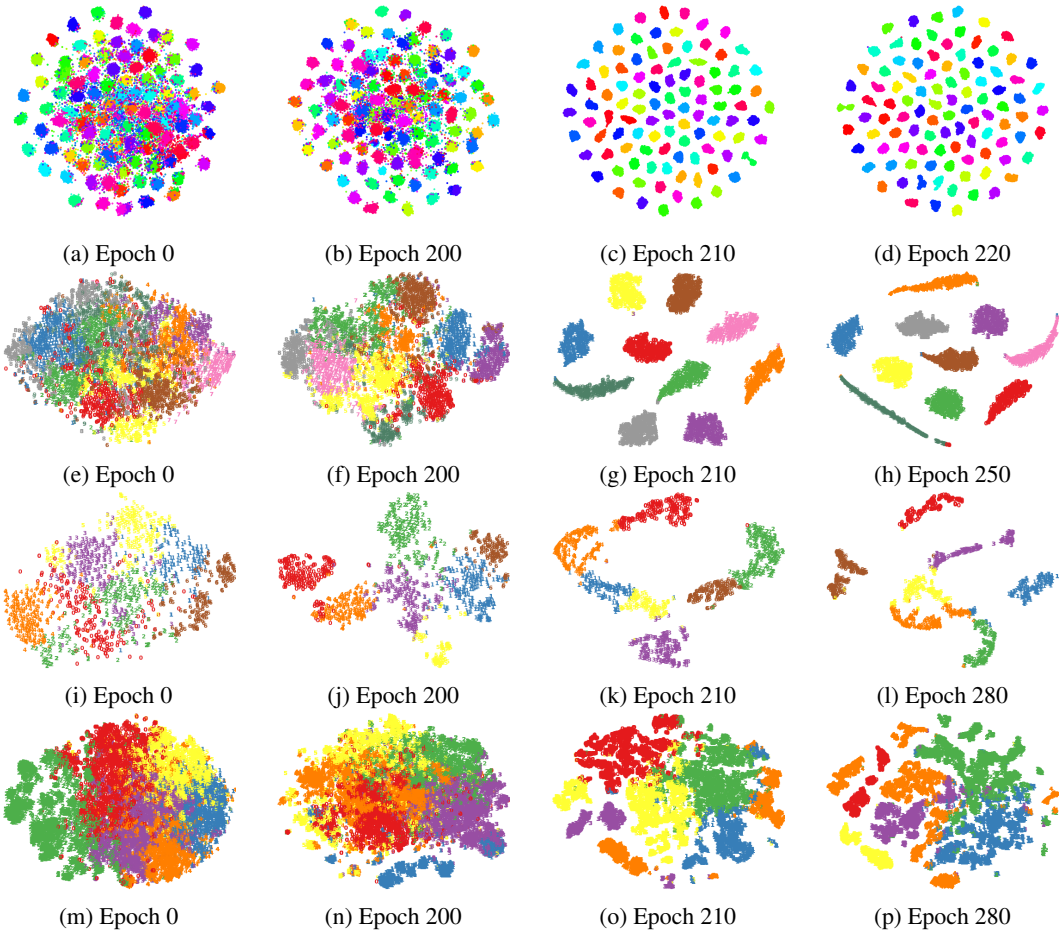
We further analyze the sensitivity of key hyperparameters on additional datasets. Figure 10 presents the ACC and PUR results for Caltech-5V and NGs, which are not reported in the main text. Figure 8 shows the effect of varying the loss weights λ_1 and λ_2 in Eq. (15) on other datasets. Figure 9 presents the effect of the temperature parameters τ_1 and τ_2 in Eq. (4) and Eq. (14) on other datasets. Small temperatures lead to larger fluctuations in clustering performance, while moderate to high values result in stable performance, confirming that the model is relatively insensitive to the choice of τ_1 and τ_2. Overall, these additional visualizations support the hyperparameter choices reported in the main text and demonstrate that VIR-SCI maintains robust performance across reasonable ranges of loss weights and temperature parameters.

A.3.5 Additional Ablation Study

We further report the ablation study results on the MNIST-USPS and Fashion datasets in Table 4. The trends observed on these datasets are consistent with those reported in the main text. Specifically, relying only on L_{rec} (M_1) results in very low clustering metrics on both datasets. Using only L_{MI} (M_2) or $L_{\text{cl}}^{\mathbf{R}^v}$ (M_3) significantly improves the metrics, but they still fail to reach optimal performance. Comparisons of M_3 vs. M_4, M_6 vs. M_7, and M_8 vs. M_9 show that simultaneously applying symmetric

cross-view incremental reasoning, feature-wise relevance modulation, and the additive fusion strategy further enhances clustering performance.

Removing the symmetric cross-view incremental reasoning (M_{20}) causes a performance decrease (ACC decreases by 0.11% on Fashion), indicating that iterative view-wise reasoning effectively exploits complementary information. Without feature-wise relevance modulation (M_{18}), redundant or irrelevant features interfere with the current view, resulting in an ACC drop of 0.12% on MNIST-USPS. Excluding the current view during fusion (M_{16}) further reduces performance on Fashion, highlighting the importance of maintaining view-specific intrinsic semantics. Sequentially removing these modules results in M_{10}, M_{12}, and M_{14}, which also confirms the effectiveness of each component. The relatively small performance variations on these two datasets are because introducing any single component already yields a large improvement, approaching saturation. That the model still benefits from additional components under such conditions demonstrates that the proposed modules collaborate effectively to learn higher-quality multi-view representations.



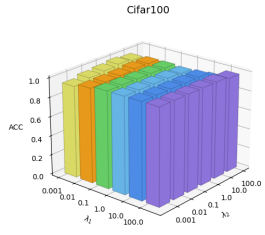
(a) Epoch 0 (b) Epoch 200 (c) Epoch 210 (d) Epoch 220

(e) Epoch 0 (f) Epoch 200 (g) Epoch 210 (h) Epoch 250

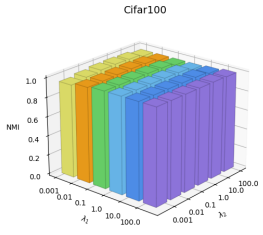
(i) Epoch 0 (j) Epoch 200 (k) Epoch 210 (l) Epoch 280

(m) Epoch 0 (n) Epoch 200 (o) Epoch 210 (p) Epoch 280

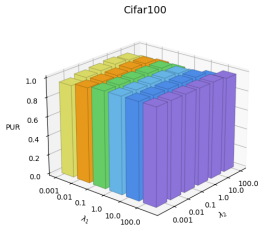
Figure 7: T-SNE visualizations of clustering results on Cifar100 ((a)–(d)), MNIST-USPS ((e)–(h)), Caltech-5V ((i)–(l)), and Reuters ((m)–(p)) datasets with increasing training epochs.



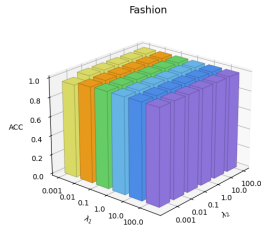
(a) λ_1 and λ_2: ACC on Cifar100



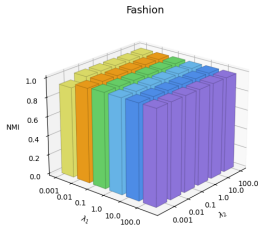
(b) λ_1 and λ_2: NMI on Cifar100



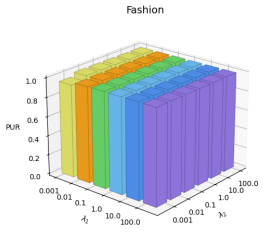
(c) λ_1 and λ_2: PUR on Cifar100



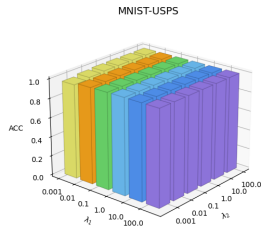
(d) λ_1 and λ_2: ACC on Fashion



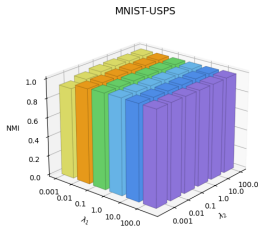
(e) λ_1 and λ_2: NMI on Fashion



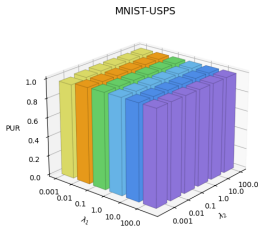
(f) λ_1 and λ_2: PUR on Fashion



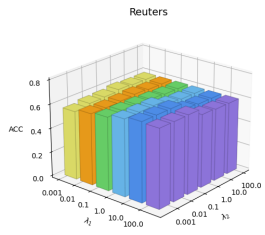
(g) λ_1 and λ_2: ACC on MNIST-USPS



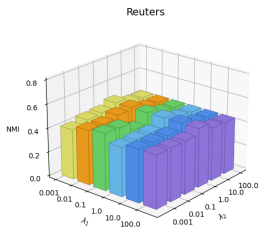
(h) λ_1 and λ_2: NMI on MNIST-USPS



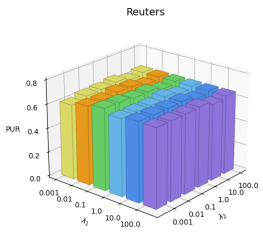
(i) λ_1 and λ_2: PUR on MNIST-USPS



(j) λ_1 and λ_2: ACC on Reuters

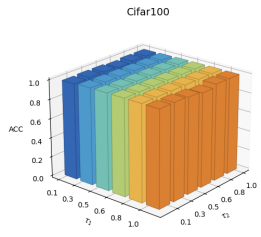


(k) λ_1 and λ_2: NMI on Reuters

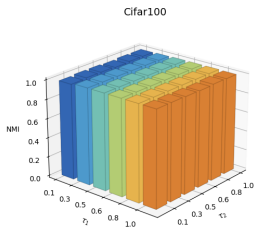


(l) λ_1 and λ_2: PUR on Reuters

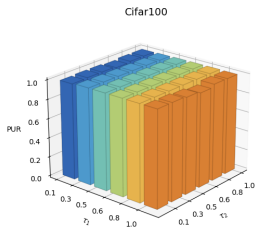
Figure 8: Parameter sensitivity analysis of λ_1 and λ_2 on Cifar100 ((a)–(c)), Fashion ((d)–(f)), MNIST-USPS ((g)–(i)), and Reuters ((j)–(l)) datasets.



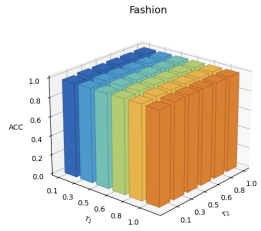
(a) τ_1 and τ_2: ACC on Cifar100



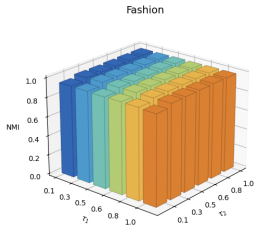
(b) τ_1 and τ_2: NMI on Cifar100



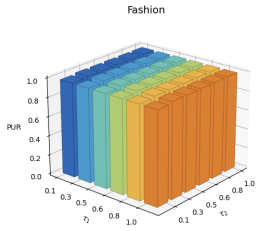
(c) τ_1 and τ_2: PUR on Cifar100



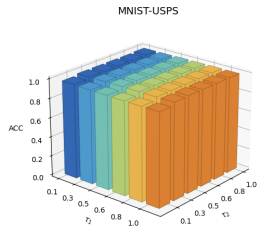
(d) τ_1 and τ_2: ACC on Fashion



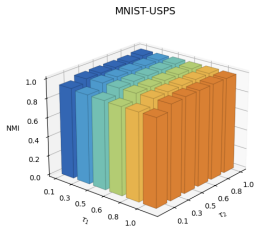
(e) τ_1 and τ_2: NMI on Fashion



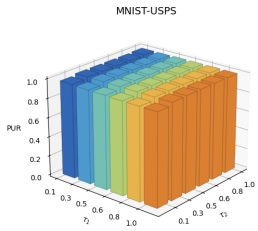
(f) τ_1 and τ_2: PUR on Fashion



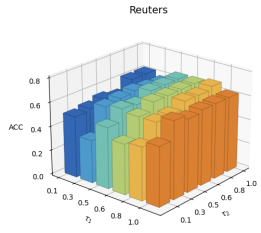
(g) τ_1 and τ_2: ACC on MNIST-USPS



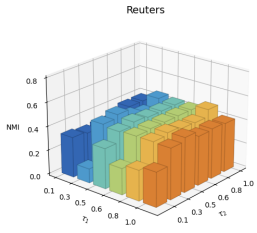
(h) τ_1 and τ_2: NMI on MNIST-USPS



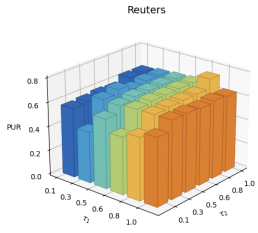
(i) τ_1 and τ_2: PUR on MNIST-USPS



(j) τ_1 and τ_2: ACC on Reuters

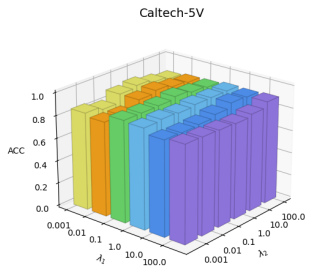


(k) τ_1 and τ_2: NMI on Reuters

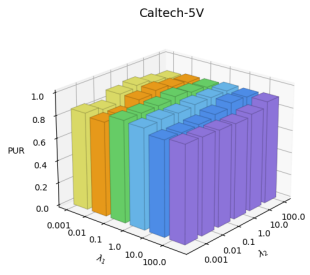


(l) τ_1 and τ_2: PUR on Reuters

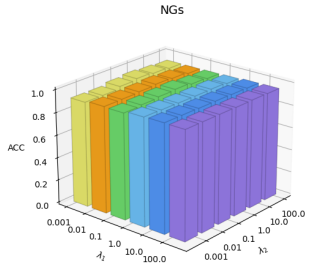
Figure 9: Parameter sensitivity analysis of τ_1 and τ_2 on Cifar100 ((a)–(c)), Fashion ((d)–(f)), MNIST-USPS ((g)–(i)), and Reuters ((j)–(l)) datasets.



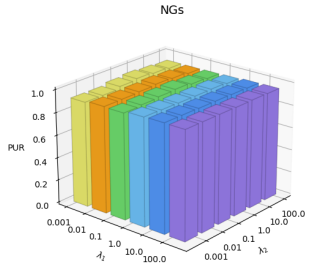
(a) λ_1 and λ_2: ACC on Caltech-5V



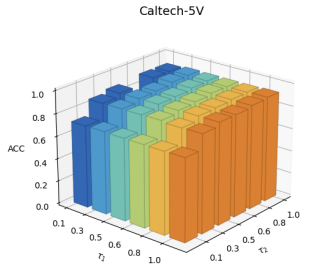
(b) λ_1 and λ_2: PUR on Caltech-5V



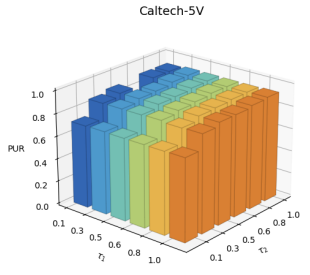
(c) λ_1 and λ_2: ACC on NGs



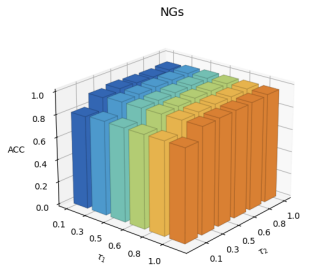
(d) λ_1 and λ_2: PUR on NGs



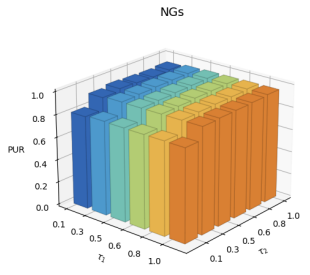
(e) τ_1 and τ_2: ACC on Caltech-5V



(f) τ_1 and τ_2: PUR on Caltech-5V



(g) τ_1 and τ_2: ACC on NGs



(h) τ_1 and τ_2: PUR on NGs

Figure 10: Parameter sensitivity results of λ_1 and λ_2: (a)-(b) show the ACC and PUR on Caltech-5V, and (c)-(d) show the ACC and PUR on NGs. Parameter sensitivity results of τ_1 and τ_2: (e)-(f) show the ACC and PUR on Caltech-5V, and (g)-(h) show the ACC and PUR on NGs.