JUSTIFIED CROSS-VIEW INTERACTION: BEYOND TRIVIAL ALIGNMENT IN MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep multi-view clustering seeks to discover coherent semantic structures from heterogeneous views by exploiting cross-view information. However, existing methods often treat cross-view interaction as inherently beneficial, overlooking that information exchanged across views can be unreliable, structurally incompatible, or redundant. To address these issues, we formulate multi-view clustering from the perspective of *Justified Cross-View Interaction* (JCVI), where an effective interaction should be both useful and non-trivial. Specifically, we develop a utility-aware interaction mechanism that jointly characterizes semantic reliability, structural compatibility, and feature complementarity, and adaptively regulates cross-view information exchange to selectively incorporate complementary cues while preserving view-specific discriminative semantics. To promote interaction non-triviality, we further construct a target-exclusive supervisory context for each view using only the remaining views, thereby removing the direct self-information path and encouraging the target representation to capture semantic structures predictable from genuinely external views. Multi-granularity structural constraints are further imposed at pairwise and neighborhood levels to preserve coherent cross-view organization. Extensive experiments on five benchmark datasets demonstrate the effectiveness of the proposed framework and its strong clustering performance against representative state-of-the-art methods.

1 INTRODUCTION

With the development of deep learning, deep multi-view clustering techniques have shown superior representation capabilities; however, existing approaches still face significant challenges in handling view-specific redundancy and achieving robust multi-level semantic alignment.

In recent years, deep multi-view clustering has achieved significant progress in mining consistent semantics from heterogeneous data by integrating auto-encoders and graph representation learning (Bian et al., 2024; Wang et al., 2026a). To improve fusion efficiency, existing studies frequently employ mechanisms such as adaptive graph learning, structural co-regularization, or sample-level weighting to dynamically suppress interference and capture subtle structures (Guo et al., 2026; Li et al., 2026b; Qiang et al., 2026; Wang et al., 2026a). In response to inter-view feature redundancy, several works have introduced information-theoretic lower bound constraints, decoupled representations, or Mixture-of-Experts (MoE) architectures to ensure that multi-view aggregation delivers genuine information gain rather than signal dilution (Xiao et al., 2026; Ren et al., 2026; Zou et al., 2026). Moreover, multi-granularity contrastive alignment has emerged as a pivotal means to enhance feature discriminability, with research covering multi-scale consistency constraints across instance, cluster, and encoding levels (Yang & Tan, 2026; Yuan et al., 2026; Cui et al., 2026; Hu et al., 2025; Bian et al., 2025; Wang et al., 2025).

Despite the substantial progress achieved, deep multi-view clustering still faces a more fundamental question: *when is cross-view interaction actually justified?* Existing methods often assume that information exchanged across views is inherently beneficial, while overlooking two distinct risks: the interaction may introduce uninformative or incompatible information, or it may exploit information that originates from the target view itself. We therefore identify two closely related challenges that limit the effectiveness of cross-view interaction.

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

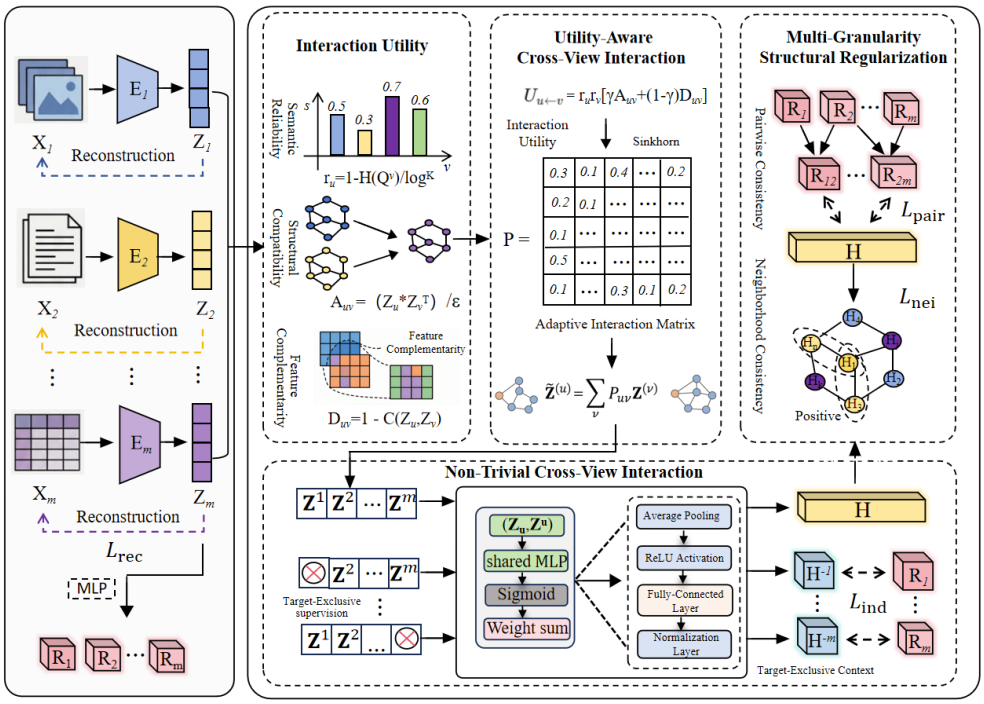


Figure 1: Overview of the proposed JCVI framework. JCVI is built on two complementary principles: (i) *interaction utility*, which jointly evaluates semantic reliability, structural compatibility, and feature complementarity to selectively regulate cross-view information exchange; and (ii) *interaction non-triviality*, which constructs a target-exclusive supervisory context to remove the direct self-information path during cross-view alignment. Pairwise- and neighborhood-level structural regularization is further imposed to preserve coherent semantic organization in the learned representation space.

The first challenge is the lack of principled criteria for determining whether cross-view interaction is worthwhile. Some existing works have recognized the presence of view bias and quality disparities across views, and have attempted to alleviate the limitation of treating all views as equally contributive through view-weighting mechanisms (Qiang et al., 2026; Li et al., 2026b). However, such approaches typically assign weights at the view level and overlook the substantial information heterogeneity among individual samples within the same view. A sample may contain highly discriminative information in one view while exhibiting task-irrelevant variations or redundant patterns in another. Other methods enhance the discriminability of anchors through anchor-to-graph structural co-regularization, aiming to exploit structural-semantic information more rigorously during fusion (Guo et al., 2026). Nevertheless, existing approaches generally lack an explicit criterion for assessing the utility of a specific cross-view interaction, i.e., whether information from one view is sufficiently reliable, structurally compatible, and informative for another. Without such adaptive evaluation, indiscriminate information exchange may introduce meaningless or redundant signals and compromise the learned semantic structure. This issue becomes increasingly pronounced as the number of views grows, where excessive cross-view interaction can weaken cluster separability and lead to suboptimal or unstable optimization.

The second challenge is the risk of trivial interaction caused by self-containing consensus supervision. To capture shared semantics across views, existing contrastive learning frameworks commonly construct a global consensus as a supervisory signal for feature alignment across branches (Zou et al., 2024; Li et al., 2026a; Bian et al., 2025). Although global consensus facilitates semantic coordination across views, the consensus often contains information originating from the target view itself. As a result, aligning a target view with such a self-containing consensus introduces a direct self-information path, allowing the model to satisfy the alignment objective without fully relying on complementary information from other views. This creates a trivial self-alignment shortcut and

weakens the incentive to discover genuinely cross-view semantic structures. Although some works (Guo et al., 2024; He et al., 2024) attempt to alleviate false negatives through contextual information or variational inference, these approaches do not explicitly remove the direct information path between a target view and its own contribution to the consensus. Consequently, successful optimization of the alignment objective does not necessarily imply effective exploitation of complementary cross-view information.

To address these fundamental challenges, we propose *JCVI* (*Justified Cross-View Interaction*), a novel deep multi-view clustering framework that rethinks cross-view learning from the perspective of whether an interaction is genuinely justified. As illustrated in Figure 1, JCVI departs from the conventional assumption that cross-view interaction is inherently beneficial and instead requires such interaction to be both useful to the target view and non-trivial with respect to the source of supervision. To model interaction utility, JCVI jointly evaluates semantic reliability, structural compatibility, and feature complementarity across views, and adaptively regulates cross-view information exchange to selectively incorporate complementary cues while preserving view-specific discriminative semantics. To enforce non-trivial interaction, JCVI constructs a target-exclusive supervisory context for each view using only the remaining views, thereby removing the direct self-information path that can lead to trivial consensus alignment and encouraging the discovery of genuinely cross-view semantic structures. Furthermore, JCVI incorporates pairwise- and neighborhood-level structural constraints to preserve coherent semantic organization after cross-view interaction, ultimately learning a unified and discriminative representation for multi-view clustering.

The main contributions of this paper are summarized as follows:

- We formulate multi-view clustering from the perspective of justified cross-view interaction, challenging the conventional assumption that cross-view interaction is inherently beneficial. We characterize effective interaction through two complementary principles: interaction utility and interaction non-triviality.
- We develop a differentiable utility-aware interaction mechanism that jointly evaluates semantic reliability, structural compatibility, and feature complementarity across views, and adaptively regulates cross-view information exchange accordingly. This enables selective incorporation of complementary information while preserving discriminative view-specific semantics.
- We enforce non-trivial cross-view interaction through an exclusion-based supervisory principle, where each target view is guided exclusively by the complementary information of the remaining views. By removing the direct self-information path, the proposed objective discourages trivial consensus alignment and promotes the discovery of genuinely cross-view semantic structures.

2 Related Work

Deep multi-view clustering aims to integrate complementary information from different sources to improve performance. We categorize existing works into three main directions.

2.1 Multi-View Semantic Fusion and Adaptive Graph Learning

This line of work integrates multi-view information while addressing interference and inconsistent view quality. Early on, MAGA (Bian et al., 2024) used self-attention for cross-view spectral graph aggregation. To capture finer structures, MCBG (Qiang et al., 2026) applied bilateral constraints on the anchor similarity matrix, while LFCGL (Wang et al., 2026a) learned sample-level importance weights for consensus graph generation. To suppress interference and enhance interactions, MLGF (Shi et al., 2026) unified graphs via distribution comparison, EGTFC (Wang et al., 2026b) employed higher-order graph trend filters, and PSGVC (Li et al., 2026a) utilized pseudo-label similarity graphs for global supervision. Additionally, to handle large-scale data, AGSCR-MVC (Guo et al., 2026) explored anchor-to-graph co-regularization, whereas NSEAN (Zhao et al., 2026) achieved one-step clustering via joint adaptive graph learning and non-negative spectral embedding.

2.2 Information Gain and Feature Redundancy Reduction

This category reduces redundant overlaps between views to produce more discriminative features. One prominent approach is representation disentanglement: VITAL (He et al., 2024) and SCDFL (Zhang & Che, 2024) decoupled features into shared/consistency and private/diversity components, which was later advanced by GMAE (Zou et al., 2026) via a dual-path autoencoder and DMCAR-MVC (Xiao et al., 2026) via a Disentangled Mixture of Experts (D-MoE). From an information-theoretic view, DDMVC (Xu et al., 2025) used decorrelation constraints for independent embeddings, while STMVC (Ren et al., 2026) constructed a sufficient representation lower bound to eliminate redundancy. Conversely, BONE (Xu et al., 2026) integrated consistency without explicit priors through high-level feature learning. Finally, ARIA (Kong et al., 2026) introduced tensor-based surrogate rank minimization for large-scale redundancy reduction.

2.3 Multi-Level and Multi-Granularity Contrastive Alignment

This direction strengthens feature structures through multi-level contrastive learning to improve consistency. For granularity exploration, TRUST (Wang et al., 2023) employed triple-granularity contrast, MCMC (Bian et al., 2025) proposed a three-level framework (instance, cluster, and prototype), 3MC (Chen et al., 2025) contrasted spaces across feature and label layers, and DMvCGSA (Fei et al., 2025) combined instance/cluster contrast with GCNs. To overcome computational and interference limitations, DWCL (Yuan et al., 2026) reduced contrastive complexity from quadratic to linear using silhouette coefficients. Meanwhile, PBDMSC (Yu et al., 2025) and STCMC-UR (Hu et al., 2025) improved robustness using pseudo-label guidance and uncertainty estimation to filter pairs. Lastly, ILASC (Yang & Tan, 2026) focused on structure alignment by matching sample neighborhood relationships.

3 Method

In this section, we present the proposed JCVI framework for deep multi-view clustering. JCVI is built on a simple principle: cross-view interaction should not be treated as universally beneficial, but should be justified by the usefulness of the exchanged information and the non-triviality of the resulting supervision. Accordingly, JCVI addresses cross-view learning from two complementary perspectives. First, it evaluates the utility of information exchanged between views by considering semantic reliability, structural compatibility, and feature complementarity, and adaptively regulates the resulting information flow. Second, it enforces non-trivial interaction by constructing the supervisory context of each target view exclusively from the remaining views, thereby removing the direct self-information path in consensus alignment. On top of these two principles, pairwise- and neighborhood-level structural constraints are introduced to preserve coherent semantic organization after cross-view interaction.

3.1 Interaction Utility Modeling

Let $\mathbf{X}^{(v)} \in \mathbb{R}^{B\times D_v}$ denote the v-th view and $\mathbf{Z}^{(v)} \in \mathbb{R}^{B\times d}$ its latent representation produced by a view-specific encoder. For each target–source pair (u, v), JCVI evaluates interaction utility from semantic reliability, structural compatibility, and feature complementarity.

We estimate the *semantic reliability* of view v from the confidence of an auxiliary clustering head:

$$\mathbf{Q}^{(v)} = \mathrm{softmax}(\mathbf{Z}^{(v)}\mathbf{C}^{\top}), \qquad r^{(v)} = 1 + \frac{1}{B\log K}\sum_{i=1}^{B}\sum_{k=1}^{K} q_{ik}^{(v)} \log q_{ik}^{(v)}, \tag{1}$$

where $\mathbf{C} \in \mathbb{R}^{K\times d}$ contains learnable cluster prototypes. A larger $r^{(v)}$ indicates more confident semantic predictions.

To measure *structural compatibility*, we compare the batch-wise similarity structures of two views:

$$\mathbf{S}^{(v)} = \mathrm{softmax!}\left(\mathbf{Z}^{(v)}\mathbf{Z}^{(v)\top}\right), \qquad A_{uv} = \exp!\left(-\frac{|\mathbf{S}^{(u)} - \mathbf{S}^{(v)}| * F^2}{\sigma^2}\right), \tag{2}$$

where σ is the bandwidth parameter; larger $A * uv$ indicates stronger structural agreement.

We further quantify *feature complementarity* as

$$D_{uv} = 1 - \frac{|\mathbf{Z}^{(u)\top}\mathbf{Z}^{(v)}|_F^2}{|\mathbf{Z}^{(u)\top}\mathbf{Z}^{(u)}|_F |\mathbf{Z}^{(v)\top}\mathbf{Z}^{(v)}| * F}, \tag{3}$$

where a larger $D * uv$ indicates less redundant and potentially more complementary representations.

The interaction utility from source v to target u is then defined as

$$U_{u\leftarrow v} = r^{(u)} r^{(v)} \left[\gamma A_{uv} + (1-\gamma) D_{uv}\right], \qquad \gamma \in (0,1). \tag{4}$$

Thus, high utility favors interactions between reliable views that are structurally compatible while retaining complementary information. The resulting $U_{u\leftarrow v}$ is used to construct the adaptive interaction mechanism in the next subsection.

3.2 Utility-Aware Cross-View Interaction

Given the pairwise interaction utilities, JCVI learns an adaptive interaction structure rather than uniformly aggregating all views. Specifically, we seek a soft interaction matrix $\mathbf{P} \in \mathbb{R}^{V\times V}$ by solving an entropy-regularized optimization problem under doubly stochastic and self-exclusion constraints:

$$\max_{\mathbf{P}} \quad \sum_{u=1}^{V}\sum_{v=1}^{V} P_{uv}U_{u\leftarrow v} - \sum_{u=1}^{V}\sum_{v=1}^{V} P_{uv}\log P_{uv} \text{ s.t.} \qquad \mathbf{P1} = \mathbf{1}, \quad \mathbf{P}^\top\mathbf{1} = \mathbf{1}, P_{vv} = 0, \forall v, \quad \mathbf{P} \geq \mathbf{0}. \tag{5}$$

The self-exclusion constraint prevents direct information exchange from a view to itself. We solve the problem using Sinkhorn–Knopp normalization. Let $\mathbf{U} = [U_{u\leftarrow v}]_{u,v}$ and define the zero-diagonal kernel

$$\mathbf{K} = \exp(\mathbf{U}) \odot (\mathbf{1}\mathbf{1}^\top - \mathbf{I}), \tag{6}$$

where the exponential is applied element-wise. Starting from $\mathbf{b}^{(0)} = \mathbf{1}$, the Sinkhorn iterations and the resulting interaction matrix are

$$\mathbf{a}^{(t)} = \mathbf{1} \oslash (\mathbf{K}\mathbf{b}^{(t-1)}), \quad \mathbf{b}^{(t)} = \mathbf{1} \oslash (\mathbf{K}^\top\mathbf{a}^{(t)}), \quad \mathbf{P} = \mathrm{diag}(\mathbf{a}^{(T)})\mathbf{K}\mathrm{diag}(\mathbf{b}^{(T)}), \tag{7}$$

where $\oslash$ denotes element-wise division.

For each target view u, the received cross-view information is aggregated as

$$\tilde{\mathbf{Z}}^{(u)} = \sum_{v=1}^{V} P_{uv}\mathbf{Z}^{(v)}, \tag{8}$$

where the learned interaction weights favor more useful source views.

To avoid over-smoothing discriminative view-specific information, we further introduce a feature-wise retention gate that considers the target representation, the aggregated representation, their absolute difference, and their element-wise product:

$$\mathbf{G}^{(u)} = \mathrm{Sigmoid}\left(\mathrm{MLP}\left(\left[\mathbf{Z}^{(u)}, \tilde{\mathbf{Z}}^{(u)}, \left|\mathbf{Z}^{(u)} - \tilde{\mathbf{Z}}^{(u)}\right|, \mathbf{Z}^{(u)} \odot \tilde{\mathbf{Z}}^{(u)}\right]\right)\right). \tag{9}$$

The refined representation is then obtained as

$$\hat{\mathbf{Z}}^{(u)} = \mathrm{LayerNorm}\left(\mathbf{G}^{(u)} \odot \mathbf{Z}^{(u)} + (1 - \mathbf{G}^{(u)}) \odot \tilde{\mathbf{Z}}^{(u)}\right). \tag{10}$$

Through this utility-aware interaction process, JCVI selectively incorporates complementary cross-view information while preserving discriminative view-specific semantics.

3.3 Non-Trivial Cross-View Interaction

Utility-aware interaction does not guarantee genuinely cross-view supervision, since a target view may still be aligned with a consensus containing its own information. To avoid such self-information recycling, JCVI constructs a target-exclusive supervisory context using only the remaining views.

Given the enhanced representations $\hat{\mathbf{Z}}^{(v)} * v = 1^V$, we first estimate the global importance of each view:

$$s^{(u)} = \mathrm{MLP}\left(\mathrm{MeanPool}\left(\hat{\mathbf{Z}}^{(u)}\right)\right) + \lambda r^{(u)}, \qquad a^{(u)} = \frac{\exp(s^{(u)})}{\sum * v = 1^V \exp(s^{(v)})}. \tag{11}$$

For each target view v, its supervisory context excludes the target itself:

$$\mathbf{H}^{(-v)} = \mathrm{MLP}\left(\sum_{u \neq v} \tilde{a}^{(u)} \hat{\mathbf{Z}}^{(u)}\right), \qquad \tilde{a}^{(u)} = \frac{a^{(u)}}{\sum_{k \neq v} a^{(k)}}. \tag{12}$$

We project each target representation into the shared embedding space as $\mathbf{R}^{(v)} = \mathrm{MLP}(\hat{\mathbf{Z}}^{(v)})$ and align it with its target-exclusive context using an instance-level contrastive objective:

$$\mathcal{L} * \mathrm{ind} = -\frac{1}{BV} \sum * i = 1^B \sum_{v=1}^{V} \log \frac{\exp\left(\mathrm{sim}\left(\mathbf{R}_i^{(v)}, \mathbf{H}^{(-v)} * i\right)/\tau\right)}{\sum * j = 1^B \exp\left(\mathrm{sim}\left(\mathbf{R}_i^{(v)}, \mathbf{H}^{(-v)} * j\right)/\tau\right)}, \tag{13}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is the temperature. By excluding the target view from its supervisory context, $\mathcal{L} * \mathrm{ind}$ encourages learning semantic structures supported by genuinely external views rather than a self-containing consensus.

3.4 Multi-Granularity Structural Regularization

To preserve semantic organization at multiple scales, we impose pairwise- and neighborhood-level structural constraints in a shared global space. We first construct the comprehensive representation

$$\mathbf{H} = \mathrm{MLP}\left(\sum_{u=1}^{V} a^{(u)} \mathbf{R}^{(u)}\right). \tag{14}$$

Pairwise Structural Consistency ($\mathcal{L}_{\mathrm{pair}}$): For each valid view pair (u, v), we construct an adaptive pairwise representation:

$$\begin{aligned} \mathbf{G}_i^{(u,v)} &= \mathrm{Sigmoid}\left(\mathrm{MLP}\left[\mathbf{R}_i^{(u)}, \mathbf{R}_i^{(v)}, |\mathbf{R}_i^{(u)} - \mathbf{R}_i^{(v)}|, \mathbf{R}_i^{(u)} \odot \mathbf{R}_i^{(v)}\right]\right), \\ \mathbf{R}_i^{(u,v)} &= \mathbf{G}_i^{(u,v)} \odot \mathbf{R} * i^{(u)} + (1 - \mathbf{G} * i^{(u,v)}) \odot \mathbf{R} * i^{(v)}. \end{aligned} \tag{15}$$

We align each pairwise representation with the comprehensive representation through

$$\mathcal{L} * \mathrm{pair} = -\frac{1}{B|\mathcal{P}|} \sum * i = 1^B \sum * (u, v) \in \mathcal{P} \log \frac{\exp\left(\mathrm{sim}(\mathbf{R}_i^{(u,v)}, \mathbf{H} * i)/\tau\right)}{\sum * j = 1^B \exp\left(\mathrm{sim}(\mathbf{R}_i^{(u,v)}, \mathbf{H}_j)/\tau\right)}, \tag{16}$$

where $\mathcal{P}$ denotes the set of valid view pairs.

Neighborhood Structural Consistency ($\mathcal{L} * \mathrm{nei}$): We construct view-specific similarity graphs and aggregate them into a reliability-weighted consensus graph:

$$\mathbf{S}^{(v)} = \mathrm{softmax}\left(\mathbf{R}^{(v)} \mathbf{R}^{(v)\top}\right), \qquad \mathbf{S} = \sum * v = 1^V \eta^{(v)} \mathbf{S}^{(v)}, \quad \eta^{(v)} = \frac{\exp(\bar{c}^{(v)})}{\sum_{m=1}^{V} \exp(\bar{c}^{(m)})}, \tag{17}$$

where $\bar{c}^{(v)}$ is the average entropy-based confidence of view v. For each sample i, we obtain $\mathcal{N} * i = \mathrm{TopK}(\mathbf{S} * i)$ and preserve its neighborhood structure in the comprehensive space via

$$\mathcal{L} * \mathrm{nei} = -\frac{1}{B} \sum * i = 1^B \frac{1}{|\mathcal{N} * i|} \sum * j \in \mathcal{N}_i \log \frac{\exp\left(\mathrm{sim}(\mathbf{H}_i, \mathbf{H} * j)/\tau\right)}{\sum * m \neq i \exp\left(\mathrm{sim}(\mathbf{H} * i, \mathbf{H} * m)/\tau\right)}. \tag{18}$$

Together, $\mathcal{L} * \mathrm{pair}$ and $\mathcal{L} * \mathrm{nei}$ preserve complementary pairwise and neighborhood structures.

3.5 Overall Objective

The overall objective jointly optimizes representation reconstruction, non-trivial cross-view supervision, and multi-granularity structural consistency. To retain the foundational data distributions of the original views, we maintain a standard autoencoder reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \sum_{v=1}^{V} \left\| \mathbf{X}^{(v)} - \hat{\mathbf{X}}^{(v)} \right\|_F^2 . \tag{19}$$

The complete training objective is formulated as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \lambda_1 \mathcal{L}_{\text{ind}} + \lambda_2 \mathcal{L}_{\text{pair}} + \lambda_3 \mathcal{L}_{\text{nei}}, \tag{20}$$

where λ_1, λ_2, and λ_3 balance the non-trivial interaction objective and the two structural regularization terms against the reconstruction objective. Upon convergence, standard k-means is applied to the learned comprehensive representation $\mathbf{H}$ to obtain the final clustering assignments. The overall training procedure is provided in Algorithm 1 in the Appendix.

4 Experiments

4.1 Experimental Setup

To evaluate the effectiveness and robustness of our method, we conduct experiments on five diverse multi-view benchmark datasets. Specifically, Fashion (Xu et al., 2021) and Synthetic3D (Kumar et al., 2011) are employed as standard visual benchmarks capturing different feature descriptors and synthetic perspectives. To assess cross-domain clustering, we utilize the MNIST-USPS (Savaram et al., 2024) handwritten digit dataset. We also incorporate Caltech-5V (Fei-Fei et al., 2004), which constructs multiple views via distinct feature extraction techniques. Finally, the BDGP (Cai et al., 2010) dataset is included to demonstrate the model's capability on real-world heterogeneous data comprising both visual and textual modalities.

To validate the superiority of our proposed framework, we benchmark it against twelve state-of-the-art deep multi-view clustering methods. Specifically, for structural and alignment-based approaches, we include CSOT (Zhang et al., 2024), which utilizes optimal transport for cross-view semantic matching, DealMVC (Yang et al., 2023), which extracts discriminative patterns through pseudo-label graph learning, and MISCC (Li et al., 2026b), which facilitates intra-view discriminability and inter-view complementarity through cross-view correlation and centralized clustering alignment. Furthermore, we consider a series of contrastive learning paradigms: SCM (Luo et al., 2024) couples contrastive objectives with graph alignment, SEM (Xu et al., 2023) focuses on semantic preservation, and CVCL (Chen et al., 2023) extends these principles for large-scale scalability. Notably, the recent HCLAA (Ding et al., 2026) performs subtle semantic decoupling via a hierarchical contrastive architecture equipped with a dual-path attention mechanism. Regarding feature representation and fusion, our baselines encompass DDMVC (Xu et al., 2025) for discriminative latent learning, and MFLVC (Xu et al., 2022), which provides a unified architecture for joint feature extraction and fusion. Finally, we evaluate against diverse advanced strategies: MAGA (Bian et al., 2024) models inter-view dependencies via attention mechanisms, SSLNMVC (Yan et al., 2025) enforces consensus consistency using KL divergence, and GFSAF (Liu et al., 2026) combats representation degradation by unifying granular-ball learning for complementary features with multi-head cross-view attention fusion.

4.2 Comparative Experimental Results Analysis

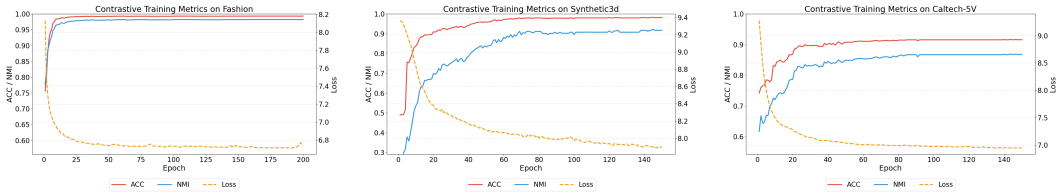
To evaluate the effectiveness and generalization of JCVI, we conduct extensive experiments on representative multi-view benchmarks. As reported in Table 1, JCVI consistently achieves the best overall performance across most metrics (ACC, NMI, and PUR). Specifically, it attains an ACC of 99.24% on the heterogeneous BDGP dataset, and exceeds 99.4% ACC and 98.5% NMI on visual datasets like Fashion and MNIST-USPS. These strong results validate the proposed justified cross-view interaction principle. By combining target-exclusive supervision that prevents direct self-information recycling with multi-granularity structural regularization, JCVI selectively integrates reliable complementary information while preserving coherent semantic structures, resulting in robust and discriminative clustering representations.

Table 1: Comparison of clustering performance (ACC, NMI, PUR) of different methods on five multi-view datasets.

Method	Fashion			Synthetic3d			MNIST-USPS			Caltech-5V			BDGP		
	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
MFLVC (2022)	99.16	97.95	99.16	95.67	83.67	95.67	99.60	98.84	99.60	85.36	75.58	85.36	98.00	94.79	98.00
CVCL (2023)	99.31	98.20	99.31	89.20	67.17	89.20	98.80	95.97	97.02	80.24	72.36	80.24	97.26	95.48	97.26
CSOT (2024)	99.14	97.86	99.14	96.83	86.60	96.83	99.42	98.34	99.42	87.21	79.13	87.21	97.88	94.19	97.88
DealMVC (2023)	82.07	91.44	82.63	80.67	56.54	80.67	87.06	92.70	89.16	86.86	78.89	86.86	78.16	84.08	78.16
SCM (2024)	81.86	78.16	82.03	94.50	79.86	94.50	98.82	96.74	98.82	74.79	65.67	74.79	75.04	71.13	75.04
DDMVC (2025)	77.17	77.59	77.48	93.47	82.56	93.47	89.30	90.94	89.30	73.79	66.39	73.79	75.28	67.36	75.28
MAGA (2024)	98.85	97.26	98.85	95.50	82.16	95.50	99.24	97.78	99.24	70.14	66.51	73.29	85.72	75.66	85.72
SEM (2023)	99.29	98.19	99.29	95.83	82.91	95.83	91.74	88.63	91.74	87.14	79.25	87.14	77.16	64.74	77.16
SSLNMVC (2025)	99.27	98.18	99.27	95.83	84.52	95.83	99.24	97.87	99.24	87.36	77.86	87.36	98.20	94.86	98.20
HCLAA (2026)	99.37	98.35	99.37	97.00	87.01	97.00	99.60	98.82	99.60	90.20	84.31	90.20	99.16	96.91	99.16
MISCC (2026)	83.94	92.46	83.94	80.33	68.73	80.33	88.54	93.41	88.54	87.46	79.63	87.46	78.36	80.08	78.36
GFSAF (2026)	99.17	98.00	99.17	97.54	89.21	97.54	99.56	98.78	99.56	89.94	82.42	89.94	99.06	96.54	99.06
JCVI	**99.47**	**98.56**	**99.47**	**97.67**	**90.79**	**97.67**	**99.74**	**99.26**	**99.74**	**90.71**	**85.33**	**90.71**	**99.24**	**97.19**	**99.24**

4.3 Model Analysis

4.3.1 Convergence Analysis



(a) Fashion (b) Synthetic3D (c) Caltech-5V

Figure 2: Evaluation metrics as a function of training epochs on Fashion, Synthetic3D, and Caltech-5V datasets.

Figure 2 illustrates the convergence behavior of JCVI on three datasets: Fashion, Synthetic3D, and Caltech-5V. The training loss consistently decreases during optimization, while the clustering metrics, including ACC and NMI, rapidly improve and gradually stabilize across the three datasets. These results indicate that JCVI can be optimized effectively and reaches stable clustering performance within a relatively small number of training epochs. The observed convergence behavior is consistent with the proposed justified cross-view interaction principle, where utility-aware information exchange facilitates the selective incorporation of complementary cues, while target-exclusive supervision reduces the opportunity for direct self-information recycling during alignment. Overall, the results provide empirical evidence of stable optimization and effective representation learning across diverse multi-view scenarios.

4.3.2 Parameter Sensitivity Analysis

Figure 3 illustrates the clustering performance of JCVI under different temperature coefficients $\tau \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$. The temperature coefficient controls the scale of similarity scores in the contrastive objectives, including $\mathcal{L}_{\text{ind}}$, $\mathcal{L}_{\text{pair}}$, and $\mathcal{L}_{\text{nei}}$. Across the five datasets, JCVI maintains consistently strong clustering performance over the tested range of τ, with only moderate variations in ACC, NMI, and PUR. In particular, the performance remains relatively stable at both lower and higher temperature values, suggesting that JCVI is not overly sensitive to the choice of τ within the evaluated range. These results indicate that the combination of target-exclusive supervision and multi-granularity structural regularization provides a relatively robust optimization behavior with respect to the similarity scale.

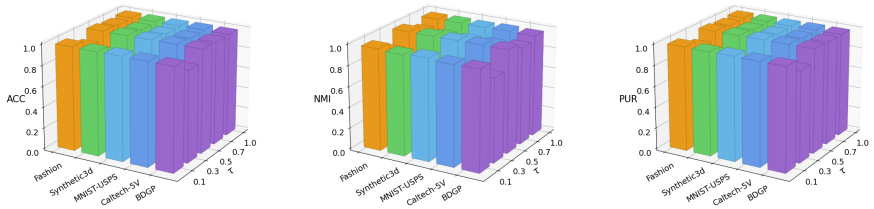


Figure 3: Analysis of clustering performance under varying temperature coefficients with ACC, NMI, and PUR metrics.

Table 2: Ablation study of different modules on five datasets.

Modules	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{ind}}$	$\mathcal{L}_{\text{pair}}$	$\mathcal{L}_{\text{nei}}$	Fashion			Synthetic3d			MNIST-USPS			Caltech-5V			BDGP		
					ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
M1	✓				62.64	58.52	63.25	49.83	25.67	57.83	43.48	30.55	43.48	79.14	73.02	79.14	62.44	39.45	62.44
M2	✓	✓			94.42	92.67	94.42	60.33	41.84	63.83	99.36	98.21	99.36	84.93	73.52	84.93	96.32	89.46	96.32
M3	✓		✓		84.32	79.18	84.32	55.50	32.89	60.83	54.81	77.33	58.19	83.21	70.57	83.21	79.84	64.36	79.84
M4	✓			✓	62.82	57.42	62.82	81.33	46.88	81.33	44.78	36.46	44.78	80.14	68.84	82.07	93.00	84.85	93.00
M5	✓	✓	✓		99.34	98.47	99.34	85.00	66.72	85.00	99.68	99.06	99.68	85.86	76.30	85.86	98.28	95.21	98.28
M6	✓		✓	✓	82.31	84.06	82.31	89.17	66.98	89.17	65.84	52.37	65.84	85.71	76.24	85.71	95.36	87.75	95.36
M7	✓	✓		✓	98.19	95.97	98.19	95.17	81.45	95.17	99.42	98.36	99.42	87.64	80.96	87.64	99.00	97.01	99.00
M8	✓	✓	✓	✓	**99.47**	**98.56**	**99.47**	**97.67**	**90.79**	**97.67**	**99.74**	**99.26**	**99.74**	**90.71**	**85.33**	**90.71**	**99.24**	**97.19**	**99.24**

4.4 Ablation Experiments

Table 2 reports the ablation results of JCVI on five benchmark datasets. The baseline model (M1) with only reconstruction loss ($\mathcal{L}_{\text{rec}}$) yields suboptimal performance, indicating its insufficiency in capturing cross-view semantic relationships. Introducing the non-trivial interaction objective $\mathcal{L}_{\text{ind}}$ (M2) leads to pronounced improvements (e.g., ACC increases from 62.64% to 94.42% on Fashion and 43.48% to 99.36% on MNIST-USPS), highlighting the benefit of target-exclusive supervision for semantic correspondence. Meanwhile, adding pairwise ($\mathcal{L}_{\text{pair}}$, M3) or neighborhood ($\mathcal{L}_{\text{nei}}$, M4) structural regularization individually improves the baseline by capturing complementary structural information. The complete JCVI configuration (M8), which jointly incorporates $\mathcal{L}_{\text{ind}}$, $\mathcal{L}_{\text{pair}}$, and $\mathcal{L}_{\text{nei}}$, achieves the best overall performance. These results demonstrate that the non-trivial cross-view interaction effectively aligns semantic correspondence across views, while the multi-scale regularizations further preserve the underlying structural organization.

5 Conclusion

In this paper, we propose JCVI, a deep multi-view clustering framework that revisits cross-view learning from the perspective of justified interaction. JCVI models interaction utility according to semantic reliability, structural compatibility, and feature complementarity, enabling the model to selectively incorporate informative cross-view cues instead of treating all interactions equally. To further ensure that the learned supervision is genuinely cross-view, target-exclusive contexts are constructed to prevent direct self-information recycling during representation alignment. In addition, pairwise- and neighborhood-level structural regularization is introduced to preserve semantic consistency at multiple granularities. Extensive experiments on five benchmark datasets demonstrate that JCVI achieves strong clustering performance and consistently benefits from its key components. These results support the effectiveness of jointly considering interaction usefulness, interaction non-triviality, and structural organization in deep multi-view clustering. Nevertheless, the current formulation assumes a fixed set of available views and has not been specifically designed for incomplete or dynamically varying multi-view settings. Extending JCVI to more flexible and challenging multi-view scenarios will be an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and proofreading. We have not used generative AI tools for generating core research ideas or writing original methodology, and generating figures or experimental data are not applicable to this work. Additionally, we did not use generative AI tools for any other tasks. We have reviewed all AI-assisted work. Specifically, all AI-edited text was manually read and verified by the authors to ensure scientific accuracy and preserve the original intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work primarily focuses on foundational machine learning methodology and does not involve human subjects, private datasets, or present foreseeable societal harms. Therefore, we do not foresee any ethical risks associated with this research.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we have provided comprehensive details regarding the experimental setups, including hyperparameter configurations and parameter experiments, in Appendix A.2.1 and Appendix A.2.3. Furthermore, we have uploaded our complete source code in the supplementary materials to facilitate replication. All datasets used in our experiments are publicly available and have been properly cited in the main text and references.

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.

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.

Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1548–1560, 2010.

Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16752–16761, 2023.

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, Xiaohuang Wu, Haitao Zhang, Chongjie Dong, and Jie Wen. Structure-guided deep multi-view clustering. *Information Fusion*, 125:103461, 2026.

Xiaojian Ding, Xian Li, Lin Zhao, and Xiaoying Zhu. Hclaa: Hierarchical contrastive learning with adaptive attention. *Information Sciences*, 741:123249, 2026.

Lunke Fei, Junlin He, Qi Zhu, Shuping Zhao, Jie Wen, and Yong Xu. Deep multi-view contrastive clustering via graph structure awareness. *IEEE Transactions on Image Processing*, 34:3805–3816, 2025. doi: 10.1109/TIP.2025.3573501.

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.

Jipeng Guo, Xiang Xu, Yu Cao, Man Cao, Mengyuan Xin, Tianxiang Zhao, Ye Su, Junbin Gao, Mingliang Cui, and Youqing Wang. Anchor-to-graph structural co-regularization for scalable multi-view clustering. *Pattern Recognition*, pp. 113378, 2026.

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.

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.

Zisen Kong, Dongxia Chang, Yiming Wang, Pengyuan Li, and Yao Zhao. Tensorial multi-view clustering via alternative rank minimization and inter-view alignment. *IEEE Transactions on Knowledge and Data Engineering*, 38(5):3196–3209, 2026. doi: 10.1109/TKDE.2026.3666476.

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

Guojie Li, Zhiwen Yu, Kaixiang Yang, Jianming Lv, and C. L. Philip Chen. Pseudo-label similarity graph-driven multi-view contrastive clustering. *IEEE Transactions on Multimedia*, 28: 4786–4798, 2026a. doi: 10.1109/TMM.2026.3660170.

Pengyuan Li, Dongxia Chang, Yiming Wang, Man Liu, Zisen Kong, Linhua Kong, and Yao Zhao. Deep multi-view clustering with intra-view similarity and cross-view correlation learning. *IEEE Transactions on Multimedia*, 28:3682–3693, 2026b. doi: 10.1109/TMM.2026.3654390.

Shuaiyu Liu, Song Wu, Jie Xu, Yazhou Ren, Yang Yang, Xiaorong Pu, and Guoying Wang. Views attention fusion of granular-ball fuzzy representations split for improved multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23810–23818, 2026.

Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. Simple contrastive multi-view clustering with data-level fusion. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 4697–4705, 2024.

Qianyao Qiang, Bin Zhang, Yunjia Hua, and Feiping Nie. Multi-view clustering via bilaterally constrained anchor graph. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48 (8):9064–9077, 2026. doi: 10.1109/TPAMI.2026.3678628.

Yazhou Ren, Zichen Wen, Junlong Ke, Chenhang Cui, Yonghao Huang, Xinyue Chen, Philip S. Yu, and Lifang He. Enhancing multi-view clustering: A sufficient information-theoretic approach for consistency acquisition and redundancy elimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):7420–7435, 2026. doi: 10.1109/TPAMI.2026.3665701.

Padmaja Savaram, Sasidhar Bandu, et al. Comparative analysis of deep learning algorithms for handwritten digit recognition. In *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)*, pp. 1–5. IEEE, 2024.

Chenhui Shi, Yongjie Xin, Haifeng Yang, Jianghui Cai, Jie Wang, Lichan Zhou, Yanting He, Fuxing Cui, Xujun Zhao, and Yaling Xun. Multi-view clustering based on the association of graph structure and feature distribution. *Information Processing & Management*, 63(4):104586, 2026.

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

Jing Wang, Songhe Feng, Gengyu Lyu, and Zhibin Gu. Triple-granularity contrastive learning for deep multi-view subspace clustering. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 2994–3002, 2023.

Penglei Wang, Jitao Lu, Danyang Wu, Rong Wang, and Feiping Nie. Enhance before fusion: Multi-view graph clustering with graph trend filter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6297–6313, 2026b. doi: 10.1109/TPAMI.2026.3655829.

Qianqian Wang, Zihao Zhang, Wei Feng, Zhiqiang Tao, and Quanxue Gao. Contrastive multi-view subspace clustering via tensor transformers autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21207–21215, 2025.

Baili Xiao, Ke Liang, Jiaqi Jin, Jun Wang, Yinbo Xu, Siwei Wang, and En Zhu. Dmcar: Disentangled mixture-of-experts with context-aware routing for multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27055–27063, 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, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9234–9243, 2021.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16051–16060, 2022.

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, 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. doi: 10.1109/TMM.2025.3543075.

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

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

Yongbo Yu, Zhoumin Lu, Feiping Nie, Weizhong Yu, Zongcheng Miao, and Xuelong Li. Pseudo-label guided bidirectional discriminative deep multi-view subspace clustering. *IEEE Transactions on Knowledge and Data Engineering*, 37(7):4213–4224, 2025. doi: 10.1109/TKDE.2025.3562723.

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. doi: 10.1109/LSP.2024.3408606.

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.

Xin Zou, Ruimeng Liu, Chang Tang, Zhenglai Li, Xinwang Liu, Kunlun He, and Wanqing Li. Learning disentangled representations for generalized multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(9):10489–10503, 2026. doi: 10.1109/TPAMI.2026.3687339.

A Appendix

A.1 Methodological Supplement

A.1.1 Illustration of Justified Cross-View Interaction

To provide a more intuitive understanding of the core mechanism in our framework, we visualize the justified cross-view interaction process. As depicted in Figure 4, this mechanism dynamically evaluates the interaction utility between any pair of views. Rather than blindly fusing information, it strictly requires the interaction to satisfy three complementary criteria: semantic reliability, structural compatibility, and feature complementarity. This rigorous filtering prevents the integration of redundant or harmful noise, ensuring that only highly beneficial cross-view cues are leveraged for representation learning.

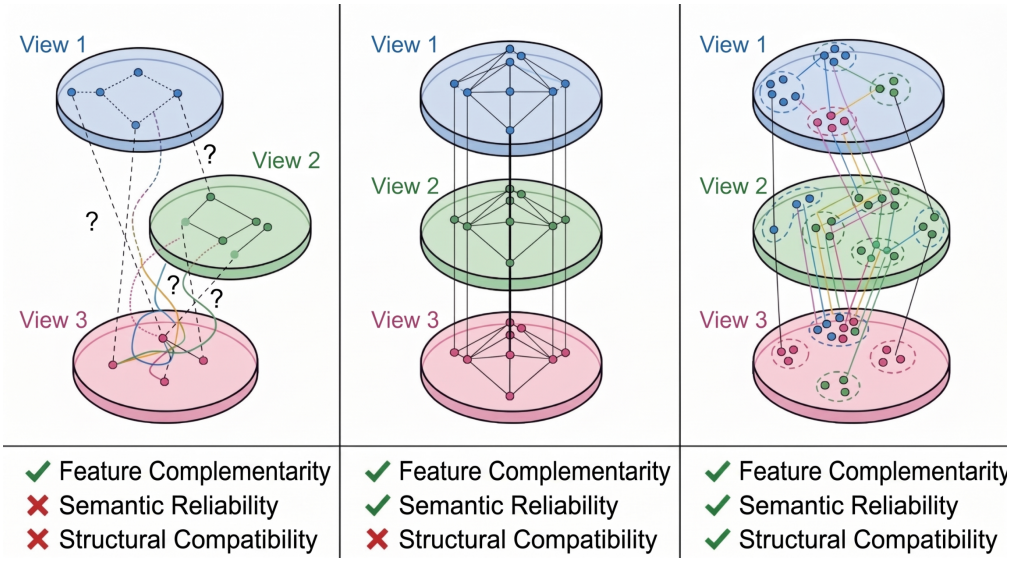


Figure 4: Illustration of justified cross-view interaction. The interaction utility between two views is evaluated from three complementary aspects: semantic reliability, structural compatibility, and feature complementarity. Only when all three criteria are satisfied is the cross-view interaction considered justified, enabling the model to selectively exploit complementary information while avoiding unreliable, structurally incompatible, or redundant information.

A.1.2 Overall Training Algorithm

In addition to the conceptual illustration, we provide the complete optimization workflow of our proposed framework. Algorithm 1 outlines the step-by-step training procedure. At each epoch, the model first extracts view-specific features and evaluates the cross-view utilities. It then constructs the target-exclusive supervisory contexts and computes the multi-granularity structural consistency losses ($\mathcal{L}_{\text{pair}}$ and $\mathcal{L}_{\text{nei}}$) alongside the non-trivial interaction loss ($\mathcal{L}_{\text{ind}}$). Finally, the network parameters are updated end-to-end, and standard k-means is applied to the optimized comprehensive representation to obtain the final clustering assignments.

Algorithm 1 JCVI Training Procedure

Require: Multi-view data $\mathbf{X}^{(v)}{}_{v=1}^{V}$, number of clusters K, training epochs E, trade-off parameters $\lambda_1, \lambda_2, \lambda_3$
Ensure: Comprehensive representation $\mathbf{H}$ and clustering assignments $\mathbf{y}$
for $epoch = 1$ **to** E **do**
  Encode each view into $\mathbf{Z}^{(v)}$ and compute $\mathcal{L}_{\text{rec}}$
  Compute semantic reliability $r^{(v)}$, structural compatibility A_{uv}, feature complementarity D_{uv}, and interaction utility $U_{u \leftarrow v}$
  Construct the zero-diagonal kernel $\mathbf{K}$ and obtain the interaction matrix $\mathbf{P}$ via Sinkhorn normalization
  Aggregate cross-view information and obtain enhanced representations $\hat{\mathbf{Z}}^{(v)}$
  Construct target-exclusive contexts $\mathbf{H}^{(-v)}$ and compute $\mathcal{L}_{\text{ind}}$
  Construct the comprehensive representation $\mathbf{H}$ and compute $\mathcal{L}_{\text{pair}}$ and $\mathcal{L}_{\text{nei}}$
  Update network parameters by minimizing $\mathcal{L}_{\text{total}}$ in Eq. (20)
end for
Obtain $\mathbf{y}$ by applying k-means to the optimized $\mathbf{H}$
return $\mathbf{H}$ and $\mathbf{y}$

A.2 Additional Experimental Analysis

A.2.1 Implementation Details

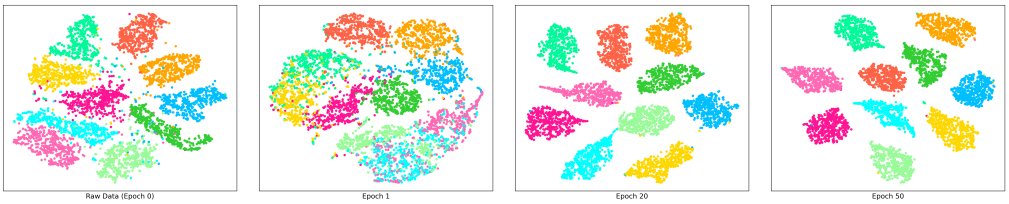
For fair comparison, inputs are vectorized and a unified configuration is applied across all datasets. The view-specific autoencoders adopt a symmetrical fully connected (FC) architecture (Input-500-500-2000-64) with hidden ReLU activations, extracting 64-dimensional latent features $\boldsymbol{Z}_v$. Subsequent MLPs project these features into view-consensus representations $\boldsymbol{R}_v$ (64-20), pairwise interaction embeddings $\boldsymbol{R}'_{ij}$ (64-20), and a unified global representation $\boldsymbol{H}$ ($M \times 64$-256-20). The network is optimized via Adam with a fixed learning rate of 3×10^{-4} and no weight decay. Experiments are implemented in PyTorch 1.10.2 on a workstation (Intel i7-12700K CPU, 32GB RAM, NVIDIA RTX 4070 GPU), with final results averaged over five independent runs to ensure reliability.

A.2.2 Clustering Results Visualization Analysis

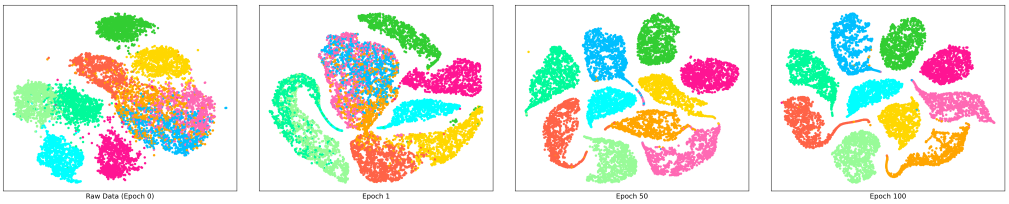
We employ t-SNE to visualize the comprehensive representations $\mathbf{H}$ learned by JCVI on the MNIST-USPS, Fashion, Caltech-5V, BDGP, and Synthetic3D datasets, as shown in Figure 5. As training progresses, the learned representations gradually exhibit clearer cluster boundaries and tighter intra-cluster distributions, particularly at later training stages. These visualizations provide qualitative evidence that JCVI progressively improves the separability and compactness of the learned representations during training. The observed evolution is consistent with the proposed justified cross-view interaction principle, which selectively incorporates complementary cross-view information while avoiding direct self-information recycling during alignment.

A.2.3 Hyperparameter Analysis

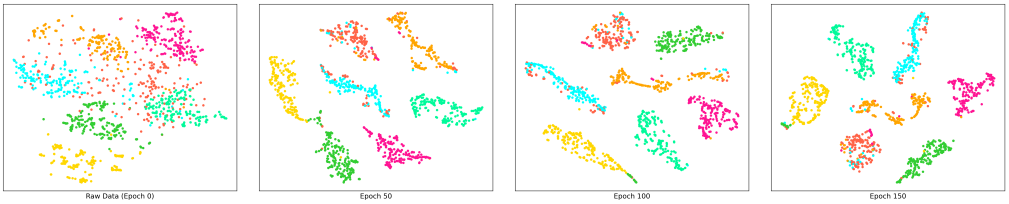
Figure 6 examines the sensitivity of JCVI to the hyperparameters λ_1, λ_2, and λ_3 on the Synthetic3D dataset, with each pair of parameters evaluated over the grid $\{0.01, 0.1, 1.0, 10.0, 100.0\}$. In the overall objective in Eq. 20, λ_1 controls the contribution of the non-trivial cross-view interaction objective $\mathcal{L}_{\text{ind}}$, while λ_2 and λ_3 control the strengths of pairwise and neighborhood structural regularization, respectively. Across the three parameter combinations, the 3D bar charts show that JCVI maintains relatively stable clustering performance over a broad range of moderate parameter values, particularly within approximately $[0.1, 10.0]$. In contrast, overly small coefficients can lead to noticeable performance degradation, suggesting that insufficient regularization weakens the contribution of non-trivial cross-view supervision and structural consistency. Excessively large coefficients may also reduce clustering performance, as overly strong auxiliary objectives can dominate the reconstruction objective and constrain the learned representation excessively. Overall, the results indicate that JCVI is relatively robust to the choice of λ_1, λ_2, and λ_3 when they are selected



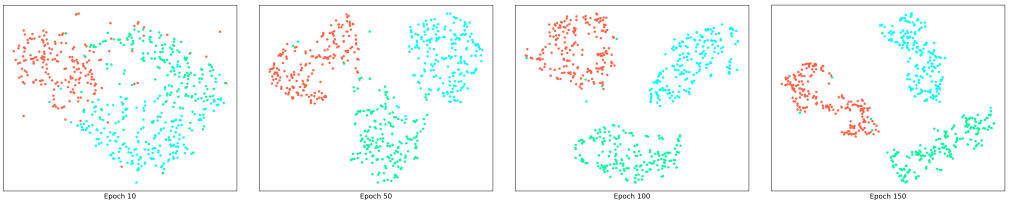
(a) Raw Data (b) Epoch 1 (c) Epoch 20 (d) Epoch 50



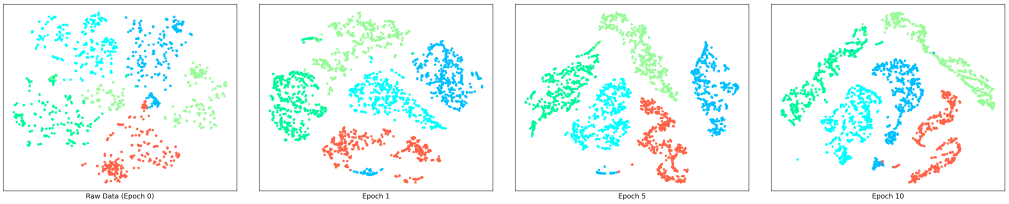
(e) Raw Data (f) Epoch 1 (g) Epoch 50 (h) Epoch 100



(i) Raw Data (j) Epoch 1 (k) Epoch 50 (l) Epoch 100



(m) Raw Data (n) Epoch 1 (o) Epoch 50 (p) Epoch 100



(q) Raw Data (r) Epoch 1 (s) Epoch 5 (t) Epoch 10

Figure 5: Clustering visualization at different training stages. From top to bottom, the rows correspond to MNIST-USPS, Fashion, Caltech-5V, Synthetic3d, and BDGP datasets, respectively. Each row displays the raw data (left) followed by epochs 1, 50, and 100, except for MNIST-USPS (epochs 1, 20, and 50) and BDGP (epochs 1, 5, and 10).

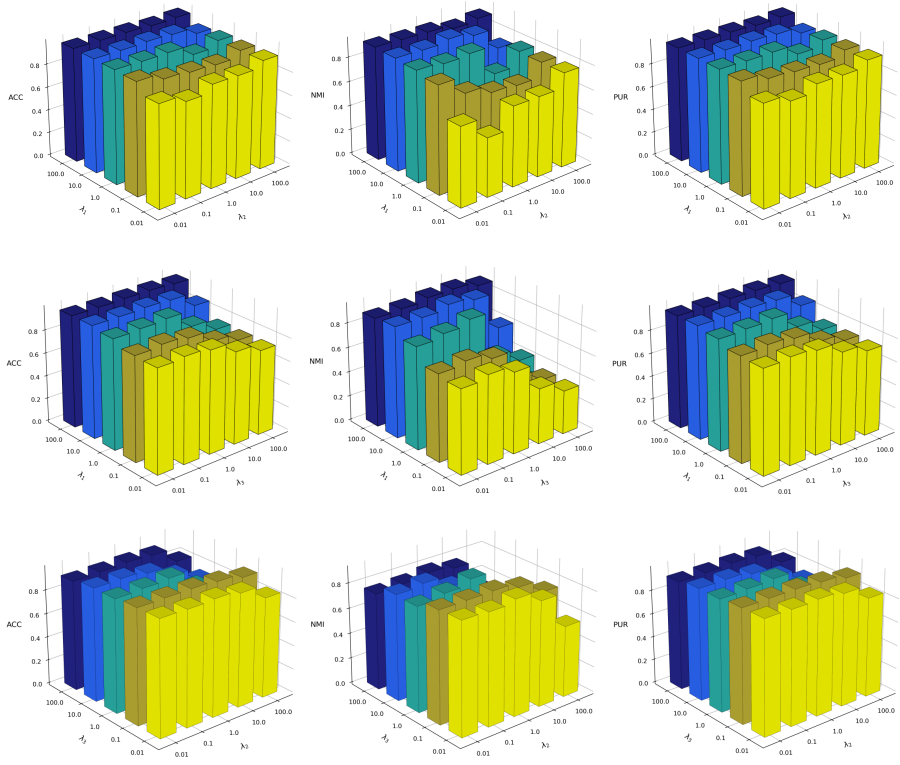


Figure 6: Analysis of clustering performance on the Synthetic3d dataset under varying hyperparameter combinations. Top row: varying λ_1 and λ_2. Middle row: varying λ_1 and λ_3. Bottom row: varying λ_2 and λ_3. The columns from left to right correspond to ACC, NMI, and PUR metrics, respectively.

within a moderate range, while a reasonable balance among non-trivial interaction and structural regularization is important for achieving strong clustering performance.

A.3 Theoretical Analysis

This section examines several mathematical properties of JCVI that clarify the behavior of its interaction utility, the role of entropy-regularized interaction learning, and the structural effect of target-exclusive supervision. These results characterize properties of the proposed mechanisms rather than providing guarantees on clustering recovery or generalization.

A.3.1 Properties of Interaction Utility

For a target-source pair (u, v), the interaction utility is

$$U_{u \leftarrow v} = r^{(u)} r^{(v)} \left[\gamma A_{uv} + (1-\gamma) D_{uv}\right], \tag{21}$$

where $r^{(u)}, r^{(v)}, A_{uv}, D_{uv} \in [0,1]$ and $\gamma \in [0,1]$. Let

$$M_{uv} = \gamma A_{uv} + (1-\gamma) D_{uv}. \tag{22}$$

Since M_{uv} is a convex combination of A_{uv} and D_{uv}, $M_{uv} \in [0,1]$, and therefore $U_{u \leftarrow v} \in [0,1]$.

Proposition 1 (Monotonicity of Interaction Utility) *For a fixed target-view reliability $r^{(u)}$, the interaction utility $U_{u \leftarrow v}$ is non-decreasing with respect to the source-view reliability $r^{(v)}$, structural compatibility A_{uv}, and feature complementarity D_{uv}.*

Proof. From Eq. (21),

$$\frac{\partial U_{u \leftarrow v}}{\partial r^{(v)}} = r^{(u)} M_{uv}, \qquad \frac{\partial U_{u \leftarrow v}}{\partial A_{uv}} = \gamma r^{(u)} r^{(v)}, \tag{23}$$

and

$$\frac{\partial U_{u\leftarrow v}}{\partial D_{uv}} = (1-\gamma) r^{(u)} r^{(v)}. \tag{24}$$

All three derivatives are non-negative under the stated assumptions. Hence, increasing any of these factors while keeping the others fixed cannot decrease the interaction utility. □

An immediate consequence is that, for a fixed target view u, if two source views v_1 and v_2 satisfy

$$r^{(v_1)} \geq r^{(v_2)}, \qquad A_{uv_1} \geq A_{uv_2}, \qquad D_{uv_1} \geq D_{uv_2}, \tag{25}$$

then

$$U_{u\leftarrow v_1} \geq U_{u\leftarrow v_2}. \tag{26}$$

Thus, a source view that is no worse under all three criteria cannot receive a lower interaction utility.

Remark on reliability attenuation. The multiplicative form in Eq. (21) directly implies that, if $\tilde{r}^{(v)} = \alpha r^{(v)}$ for $\alpha \in [0,1]$ while the other factors remain fixed, then

$$\tilde{U}_{u\leftarrow v} = \alpha U_{u\leftarrow v}. \tag{27}$$

Hence, decreasing the estimated reliability of a source view proportionally suppresses its utility, independently of its compatibility and complementarity scores.

Under the definitions used in JCVI, both A_{uv} and D_{uv} are symmetric with respect to u and v. Consequently,

$$U_{u\leftarrow v} = U_{v\leftarrow u}. \tag{28}$$

The arrow notation is retained to indicate how the resulting coefficient is used in target-wise aggregation, rather than to imply that the utility itself is asymmetric.

A.3.2 Entropy-Regularized Interaction Optimization

Given the utility matrix $\mathbf{U} = [U_{u\leftarrow v}] * u, v$, JCVI determines the interaction matrix through

$$\begin{aligned} \max_{*\mathbf{P}} \quad & \sum_{u=1}^{V}\sum_{v=1}^{V} P_{uv} U_{u\leftarrow v} - \sum_{u=1}^{V}\sum_{v=1}^{V} P_{uv}\log P_{uv} \\ \text{s.t.} \quad & \mathbf{P}\mathbf{1} = \mathbf{1}, \qquad \mathbf{P}^\top \mathbf{1} = \mathbf{1}, \qquad\qquad P_{vv} = 0, \quad \forall v, \qquad \mathbf{P} \geq \mathbf{0}. \end{aligned} \tag{29}$$

Proposition 2 (Unique Entropy-Regularized Interaction Solution) *For $V \geq 2$ and finite interaction utilities, the optimization problem in Eq. (29) admits a unique optimal interaction matrix $\mathbf{P}^*$.*

Proof. The feasible set is non-empty for $V \geq 2$; for example, any cyclic permutation matrix satisfies the non-negativity, doubly stochastic, and zero-diagonal constraints. The feasible set is also convex, closed, and bounded, and is therefore compact.

Using the convention $0 \log 0 = 0$, the objective is continuous over the feasible set. The utility term is linear in $\mathbf{P}$, whereas the entropy term

$$-\sum_{u=1}^{V}\sum_{v=1}^{V} P_{uv}\log P_{uv} \tag{30}$$

is strictly concave over the admissible entries. Therefore, the complete objective is strictly concave over the convex feasible set, which yields a unique maximizer $\mathbf{P}^*$. □

The connection between Eq. (29) and the Sinkhorn scaling used in JCVI can be seen from its optimality conditions. Define the zero-diagonal kernel

$$\mathbf{K} = \exp(\mathbf{U}) \odot (\mathbf{1}\mathbf{1}^\top - \mathbf{I}), \tag{31}$$

where the exponential is applied element-wise. Introducing Lagrange multipliers for the row- and column-sum constraints gives, for each admissible off-diagonal entry,

$$P^*_{uv} = a_u K_{uv} b_v, \qquad u \neq v, \tag{32}$$

where $a_u > 0$ and $b_v > 0$ are scaling factors. Equivalently,

$$\mathbf{P}^* = \mathrm{diag}(\mathbf{a})\mathbf{K}\mathrm{diag}(\mathbf{b}). \tag{33}$$

The vectors $\mathbf{a}$ and $\mathbf{b}$ are chosen such that the row and column sums equal one. This establishes the connection between the entropy-regularized optimization in Eq. (29) and the Sinkhorn–Knopp matrix scaling employed in Sec. 3.2.

A.3.3 EXPLICIT TARGET EXCLUSION

JCVI constructs the supervisory context of target view v using the enhanced representations of the remaining views. Recall that

$$a^{(u)} = \frac{\exp(s^{(u)})}{\sum_{m=1}^{V} \exp(s^{(m)})}, \tag{34}$$

and that the weights used for target v are re-normalized as

$$\tilde{a}^{(u)} = \frac{a^{(u)}}{\sum_{k \neq v} a^{(k)}}, \qquad u \neq v. \tag{35}$$

Substituting the softmax definition gives

$$\tilde{a}^{(u)} = \frac{\exp(s^{(u)})}{\sum_{k \neq v} \exp(s^{(k)})}, \qquad u \neq v, \tag{36}$$

so the target score $s^{(v)}$ cancels from the re-normalized weights.

The target-exclusive context can therefore be written as

$$\mathbf{H}^{(-v)} = \mathrm{MLP}\left(\sum_{u \neq v} \tilde{a}^{(u)} \hat{\mathbf{Z}}^{(u)} \right). \tag{37}$$

Proposition 3 (Explicit Target Exclusion) *Consider the context-construction mapping in Eq. (37), with the enhanced representations treated as its input variables. The target representation $\hat{\mathbf{Z}}^{(v)}$ is not an explicit input to $\mathbf{H}^{(-v)}$. Consequently,*

$$\frac{\partial \mathbf{H}^{(-v)}}{\partial \hat{\mathbf{Z}}^{(v)}} = \mathbf{0}. \tag{38}$$

Proof. Equation (36) shows that the re-normalized weights depend only on the scores of views $u \neq v$. Equation (37) likewise aggregates only $\hat{\mathbf{Z}}^{(u)}_{u \neq v}$. Thus, when the enhanced representations are regarded as the inputs to this context-construction stage, $\hat{\mathbf{Z}}^{(v)}$ does not appear in the mapping, yielding Eq. (38). □

Proposition 3 characterizes the specific role of target-exclusive supervision. A self-containing consensus may directly include the target representation in the signal used to supervise that same target, creating an identity-like shortcut. In contrast, Eq. (37) prevents such explicit self-inclusion and requires the immediate supervisory signal to be formed from the remaining enhanced views.

This result should not be interpreted as complete information isolation. Before the context in Eq. (37) is constructed, the enhanced representations $\hat{\mathbf{Z}}^{(u)}$ have already undergone the cross-view interaction described in Sec. 3.2. Therefore, an enhanced representation with $u \neq v$ may still contain information indirectly propagated from the original target view v. Accordingly, Eq. (38) establishes the absence of an explicit self-input at the context-construction stage, but does not imply statistical independence between $\mathbf{H}^{(-v)}$ and $\mathbf{X}^{(v)}$ or $\mathbf{Z}^{(v)}$.

This distinction matches the objective of JCVI: target-exclusive supervision is designed to remove the immediate self-information shortcut without preventing useful information from being exchanged across views.

A.3.4 SUMMARY OF THEORETICAL PROPERTIES

The analysis above clarifies three aspects of JCVI. First, the proposed interaction utility behaves monotonically with respect to source reliability, structural compatibility, and feature complementarity, with reliability acting as a direct attenuation factor. Second, the entropy-regularized interaction formulation admits a unique optimum and leads naturally to the matrix-scaling form used by Sinkhorn normalization. Third, target-exclusive supervision removes the explicit self-input from the supervisory-context construction stage while still permitting information exchanged through preceding cross-view interactions. These results characterize the mathematical behavior of the proposed interaction mechanisms without requiring stronger claims of clustering guarantees or statistical independence.