DISPERSION ENHANCEMENT AGAINST MSE-ALIGNMENT INDUCED REPRESENTATION COLLAPSE IN UNALIGNED MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fully unaligned multi-view clustering aims to recover unknown sample correspondences across views and uncover a shared cluster structure from unaligned representations. A common strategy is to construct learnable cross-view alignment matrices and optimize them using the mean squared error (MSE) loss for sample alignment and cross-view fusion. However, our empirical observations reveal that directly optimizing the alignment matrices and latent representations under the MSE objective can lead to a scale-shrinking shortcut, termed **MSE-alignment-induced representation collapse**, which compresses inter-sample distances and distorts the underlying cluster structure. To address this issue, we first theoretically analyze the mechanism underlying this collapse and then propose Representation Dispersion-Enhanced Unaligned Multi-view Clustering (RDE-UMC), which effectively mitigates MSE-alignment-induced representation collapse by enhancing the dispersion of latent representations while preserving discriminative cluster structures. Specifically, guided by the analyzed collapse mechanism, we first design a representation dispersion enhancement operator that quantifies dispersion across principal directions to mitigate representation collapse while maintaining the separability of representations in each view. We then incorporate this operator into the MSE objective to formulate Representation Dispersion-Enhanced MSE (RDE-MSE) to perform cross-view alignment on separable representations. Building on the aligned representations, we further integrate the operator into contrastive learning to formulate Representation Dispersion-Enhanced Contrastive Learning (RDE-CL), which further facilitates cross-view consistency fusion among the aligned representations. Experiments on benchmark datasets demonstrate that RDE-UMC effectively mitigates MSE-alignment-induced representation collapse, preserves representation dispersion and discriminability, and improves clustering.

1 INTRODUCTION

Multi-view data characterize objects through multiple sources or modalities, providing shared and complementary information across views (Chen et al., 2022; Fang et al., 2023). For instance, in personalized online recommendations, diversified individual preferences are collected from various e-commerce platforms, and comprehensively integrating these preferences is essential for reliable product recommendations. However, these views are often unaligned because different platforms store data in different orders. Conventional multi-view learning assumes known one-to-one sample correspondences; this assumption is violated in such cross-platform settings, precluding direct fusion and limiting the applicability of aligned methods (Yu et al., 2021; Ji et al., 2024; Wen et al., 2024).

The key to learning from fully unaligned multi-view data lies in fusing cross-view information despite unknown sample correspondences. A representative strategy learns a latent representation $\mathbf{Z}^v \in \mathbb{R}^{N\times d}$ for each view, constructs learnable alignment matrices $\{\mathbf{P}^{vr}\}$, and minimizes the mean squared error (MSE) between the mapped representation $\mathbf{P}^{vr}\mathbf{Z}^v$ from view v and the representation $\mathbf{Z}^r$ from view r (Yu et al., 2021; Ji et al., 2024; Zhao & Xie, 2024; Wen et al., 2024):

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{N}\sum_{v\neq r} \|\mathbf{P}^{vr}\mathbf{Z}^v - \mathbf{Z}^r\|_F^2 , \tag{1}$$

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.

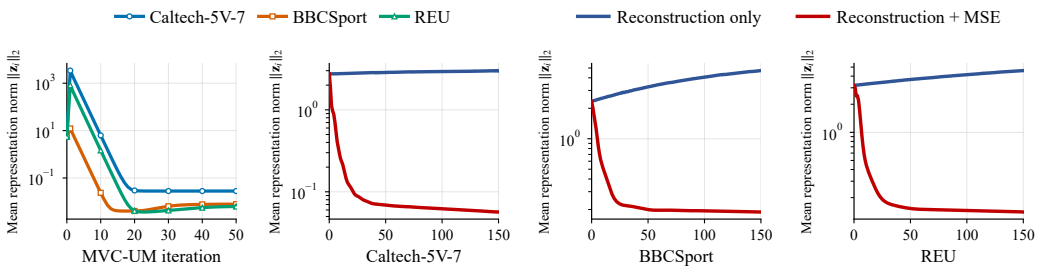


Figure 1: Preliminary evidence of representation collapse induced by the ordinary MSE alignment loss. (a) Mean sample-wise ℓ_2 representation norm, averaged across views, produced by the released MVC-UM (Yu et al., 2021) on three datasets. (b)–(d) The same statistic under autoencoder reconstruction alone and reconstruction with the ordinary MSE alignment loss $\|\mathbf{P}^{vr}\mathbf{Z}^v - \mathbf{Z}^r\|_F^2$ on Caltech-5V-7, BBCSport, and REU, respectively.

where $\mathbf{P}^{vr} \in \mathbb{R}^{N\times N}$ is a learnable non-negative cross-view mapping matrix between view v and view r, and a lower MSE is conventionally interpreted as indicating more accurate cross-view alignment. However, our experiments with the released MVC-UM implementation (Yu et al., 2021) reveal clear representation collapse. Figure 1(a) shows that optimizing the MSE alignment loss leads to a several-orders-of-magnitude reduction in the average per-sample representation norm (Haeffele et al., 2021). We further argue that although autoencoder reconstruction requires the latent representations to retain information that distinguishes individual samples, it cannot prevent the collapse caused by MSE alignment. To verify this, we conduct controlled experiments in which the same view-specific autoencoders are trained with either reconstruction alone or reconstruction combined with ordinary MSE alignment for comparison, while keeping all other settings fixed. Figures 1(b)–(d) show that reconstruction alone maintains the mean sample-wise representation norm, whereas adding the ordinary MSE alignment loss causes the mean norm to decrease consistently toward zero. These comparisons identify the ordinary MSE alignment loss as the optimization pressure that causes the observed representation collapse. We next explain this collapse tendency.

Theorem 1 (Collapse tendency of the ordinary MSE alignment loss)**.** *For each $v \neq r$, define $\mathbf{Z}^{vr} = [(\mathbf{Z}^v)^\top, (\mathbf{Z}^r)^\top]$ and $\mathbf{C}^{vr} = \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} (\mathbf{P}^{vr})^\top \\ \mathbf{0} \end{bmatrix}$. The ordinary MSE alignment loss satisfies*

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{v\neq r} \|\mathbf{Z}^{vr}\mathbf{C}^{vr} - \mathbf{Z}^{vr}\|_F^2 = \frac{1}{N}\sum_{v\neq r} \left\langle (\mathbf{Z}^{vr})^\top \mathbf{Z}^{vr}, (\mathbf{C}^{vr} - \mathbf{I})(\mathbf{C}^{vr} - \mathbf{I})^\top \right\rangle. \tag{2}$$

Minimizing the ordinary MSE alignment loss directly suppresses $(\mathbf{Z}^{vr})^\top \mathbf{Z}^{vr}$, exposing a scale-shrinking shortcut through which the objective can be reduced by driving $\mathbf{Z}^{vr}$ toward zero. As shown in Figure 1, our experiments indicate that the model tends to exploit this shortcut during optimization, resulting in representation collapse that compresses inter-sample distances and destroys the underlying cluster structure.

We further examine whether the reconstruction objective can prevent this optimization-induced collapse. During joint optimization, a sufficiently expressive decoder can compensate for arbitrarily small latent differences and preserve reconstruction accuracy even under severe collapse, provided subtle separability remains (Haeffele et al., 2021). Reconstruction loss therefore cannot prevent representation collapse. The resulting collapse weakens cross-view alignment and cluster discrimination; Appendix A.1 proves Theorem 1 and provides further analysis.

To address this failure mode, we propose representation dispersion-enhanced unaligned multi-view clustering (RDE-UMC). Geometrically, representation collapse contracts the sample distribution and shrinks the volume occupied by the representations (Ozsoy et al., 2022). Motivated by this observation, we design a representation dispersion enhancement operator that quantifies the joint representation dispersion, thereby mitigating representation collapse while maintaining their separability. We then incorporate this operator into the MSE objective to formulate Representation Dispersion-Enhanced MSE (RDE-MSE) to perform cross-view alignment on separable representations. Building on the aligned representations, we further integrate the operator into contrastive learning to formulate

Representation Dispersion-Enhanced Contrastive Learning (RDE-CL), which further facilitates cross-view consistency fusion among the aligned representations. Finally, the aligned representations are fused for clustering. Our main contributions are as follows:

- We identify **MSE-alignment-induced representation collapse** in fully unaligned multi-view clustering, theoretically characterize its scale-shrinking mechanism, and empirically validate the phenomenon through controlled experiments.
- We propose RDE-UMC with a representation dispersion enhancement operator that quantifies dispersion across principal directions to mitigate collapse. By jointly integrating this operator with MSE alignment and contrastive learning, RDE-UMC enables reliable cross-view alignment and effectively performs consistency fusion.
- Theoretical analysis characterizes the anti-collapse mechanism of RDE-UMC, while experiments on benchmark fully unaligned multi-view datasets verify its effectiveness in mitigating representation collapse and improving clustering performance.

2 Related Work

2.1 Multi-View Clustering

Multi-view clustering (MVC) recovers a shared cluster structure by modeling and fusing common and complementary information from view-specific representations (Chen et al., 2022; Fang et al., 2023). Existing methods can be broadly divided into shallow and deep models according to how they learn multi-view representations (Kumar et al., 2011; Wang et al., 2020; Liu et al., 2013; Gao et al., 2015; Zhao et al., 2017). For example, Liang et al. develop a shallow method that directly learns view-specific representations while jointly modeling consistent and view-specific structures to integrate shared and complementary information across views (Liang et al., 2022). Wang and Feng develop a deep method that combines nonlinear encoders, contrastive learning, and cross-view relation generation to extract latent representations and learn a consensus representation for clustering (Wang & Feng, 2024). Despite their effectiveness, most MVC methods assume known one-to-one sample correspondences across views. When the views are unaligned, representations at the same index may describe different objects, making direct cross-view fusion unreliable (Liu et al., 2019; Zong et al., 2020; Trosten et al., 2021).

2.2 Unaligned Multi-View Clustering

Unaligned multi-view clustering (UMC) aims to learn shared cluster structure without known cross-view sample correspondences (Yu et al., 2021; Ji et al., 2024; Zhao & Xie, 2024; Wen et al., 2024; Yu et al., 2026). Representation-level methods recover correspondences by aligning view-specific representations before fusion (Wen et al., 2024; Zhao et al., 2025). For example, MVC-UM jointly learns non-negative latent representations and cross-view mapping matrices $\mathbf{P}^{vr}$, and uses the per-view MSE alignment loss $\|\mathbf{P}^{vr}\mathbf{Z}^{v} - \mathbf{Z}^{r}\|_F^2$ to recover correspondences (Yu et al., 2021). TUMCR learns multi-scale representations, aligns them with learnable alignment matrices, and models cross-view consistency with a low-rank tensor constraint (Ji et al., 2024). However, existing studies rarely examine representation collapse caused by jointly optimizing alignment matrices and latent representations under the ordinary MSE objective, which can compromise representation discriminability and cross-view fusion (Jiang et al., 2025; Xin et al., 2026; Jiang et al., 2026).

3 Method

In this paper, we propose RDE-UMC to mitigate MSE-alignment-induced representation collapse. To counteract the resulting shrinkage in the volume occupied by latent representations, we design a representation dispersion enhancement operator that quantifies the joint representation dispersion, captures degeneration across principal directions, and maintains representation separability. We then incorporate the operator into ordinary MSE alignment to formulate RDE-MSE for cross-view alignment. After alignment, we use the same operator to adapt the contrastive temperature in RDE-CL, thereby promoting consistency fusion among the aligned representations. Finally, we fuse the aligned representations for clustering. Figure 2 illustrates the overall framework of RDE-UMC.

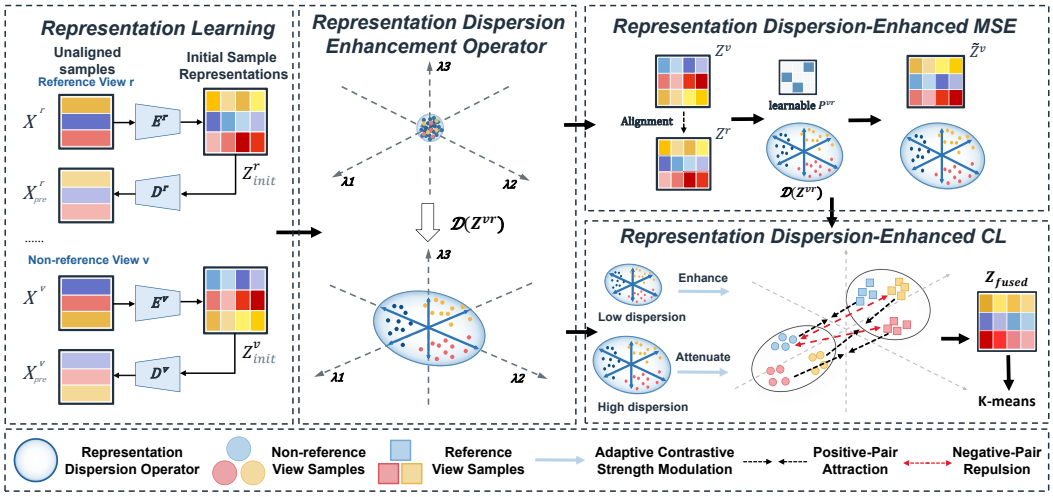


Figure 2: Overview of RDE-UMC. We first employ view-specific autoencoders to learn latent representations through reconstruction. The learned latent representations are then used to construct a representation dispersion enhancement operator that quantifies the joint representation dispersion. We next use the operator to normalize the ordinary MSE alignment loss, formulating RDE-MSE to perform cross-view alignment on separable representations. After alignment, the operator dynamically adapts the contrastive temperature in RDE-CL to promote cross-view consistency fusion. Finally, the resulting representations are fused for clustering.

3.1 Problem Definition

Consider a fully unaligned multi-view dataset $\mathcal{X} = \{\mathbf{X}^v\}_{v=1}^V$, where $\mathbf{X}^v \in \mathbb{R}^{N \times D_v}$ and $\mathbf{Z}^v \in \mathbb{R}^{N \times d}$ denote the input and latent representations of view v, respectively. Since $\mathbf{x}_i^p$ and $\mathbf{x}_i^q$ may describe different objects when $p \neq q$, direct cross-view fusion is unreliable. A common strategy selects a reference view r and learns mappings $\mathbf{P}^{vr}$ for non-reference views by minimizing Eq. (1). However, directly optimizing the alignment matrices and latent representations under the MSE objective can lead to a scale-shrinking shortcut, termed MSE-alignment-induced representation collapse, which compresses inter-sample distances and distorts cluster structure.

3.2 Representation Dispersion Enhancement Operator

To mitigate representation collapse, we design a representation dispersion enhancement operator. Specifically, we learn each view-specific latent representation using a reconstruction-based autoencoder. For view v, the encoder E^v maps $\mathbf{X}^v$ to $\mathbf{Z}^v = E^v(\mathbf{X}^v) = [\mathbf{z}_1^v, \ldots, \mathbf{z}_N^v]^\top \in \mathbb{R}^{N \times d}$, while the decoder D^v reconstructs $\widehat{\mathbf{X}}^v = D^v(\mathbf{Z}^v) \in \mathbb{R}^{N \times D_v}$. The reconstruction objective is defined as

$$\mathcal{L}_{\text{rec}} = \sum_{v=1}^{V} \frac{1}{ND_v} \left\| \mathbf{X}^v - D^v(E^v(\mathbf{X}^v)) \right\|_F^2 . \tag{3}$$

We retain $\mathcal{L}_{\text{rec}}$ during joint optimization so that the latent representations preserve view-specific information while adapting to the cross-view objectives.

Following a simple deterministic strategy, we select the view with the largest input dimension as the reference view r. The remaining views are treated as non-reference views. To mitigate MSE-alignment-induced representation collapse that tends to reduce the volume occupied by the learned representations as shown in Theorem 1, we design a representation dispersion enhancement operator that quantifies the joint representation dispersion and captures degeneration across principal directions while preserving representation separability. For each pair of representations $(\mathbf{Z}^v, \mathbf{Z}^r)$ that undergoes collapse induced by $\|\mathbf{P}^{vr}\mathbf{Z}^v - \mathbf{Z}^r\|_F^2$ as defined in Theorem 1, we form $\mathbf{Z}^{vr} = [(\mathbf{Z}^v)^\top, (\mathbf{Z}^r)^\top] \in \mathbb{R}^{d \times 2N}$ and compute its joint covariance as follows:

$$\mathbf{G}(\mathbf{Z}^{vr}) = \frac{1}{2N} \mathbf{Z}^{vr} \mathbf{H}_{2N} (\mathbf{Z}^{vr})^\top , \tag{4}$$

where $\mathbf{H}_{2N} = \mathbf{I}_{2N} - (2N)^{-1}\mathbf{1}_{2N}\mathbf{1}_{2N}^{\top}$ is the centering matrix. The covariance matrix $\mathbf{G}(\mathbf{Z}^{vr})$ describes how widely the volume occupied jointly by representations $\mathbf{Z}^v$ and $\mathbf{Z}^r$ extends. Its k-th eigenvalue $\lambda_k^{vr} \geq 0$ measures their joint spread along the corresponding principal direction of the latent space. To further mitigate collapse through this measure, we define the representation dispersion enhancement operator as

$$D(\mathbf{Z}^{vr}) = \det(\mathbf{G}(\mathbf{Z}^{vr}))^{1/d} = \left(\prod_{k=1}^{d} \lambda_k^{vr}\right)^{1/d}. \tag{5}$$

As the geometric mean of the covariance eigenvalues, $D(\mathbf{Z}^{vr})$ quantifies the joint representation dispersion of $\mathbf{Z}^v$ and $\mathbf{Z}^r$. Since both representations contribute to the joint covariance, $D(\mathbf{Z}^{vr})$ is sensitive to dispersion changes in either view. When the joint representation loses dispersion along a principal direction, the corresponding eigenvalue approaches zero and $D(\mathbf{Z}^{vr})$ decreases, allowing the operator to capture degeneration of the joint representation space.

3.3 Representation Dispersion-Enhanced MSE (RDE-MSE)

Building on the operator's dispersion-aware anti-collapse effect, we incorporate it into ordinary MSE alignment to formulate Representation Dispersion-Enhanced MSE (RDE-MSE) for reliable cross-view alignment. For each non-reference view v, we introduce a learnable non-negative mapping matrix $\mathbf{P}^{vr} \in \mathbb{R}^{N\times N}$, whose entry P_{ij}^{vr} denotes the matching degree between sample j in view v and sample i in the reference view. We then compute $D(\mathbf{Z}^{vr})$ to normalize the ordinary MSE alignment term and formulate RDE-MSE as

$$\mathcal{L}_{\text{RDE-MSE}} = \frac{1}{N}\sum_{v\neq r} \frac{\|\mathbf{P}^{vr}\mathbf{Z}^v - \mathbf{Z}^r\|_F^2}{D(\mathbf{Z}^{vr})} = \frac{1}{N}\sum_{v\neq r} \frac{\left\langle (\mathbf{Z}^{vr})^{\top}\mathbf{Z}^{vr}, (\mathbf{C}^{vr}-\mathbf{I})(\mathbf{C}^{vr}-\mathbf{I})^{\top}\right\rangle}{D(\mathbf{Z}^{vr})}. \tag{6}$$

In our design of Eq. (6), the operator is not merely a denominator that penalizes low dispersion; it cancels the quadratic scaling factor responsible for MSE-alignment-induced representation collapse. This property is summarized below.

Theorem 2 (Quadratic scaling of the representation dispersion operator)**.** *Representation collapse can be described through the scaling $\mathbf{Z}^{vr} \mapsto a\mathbf{Z}^{vr}$, where $0 < a < 1$ is a contraction factor and $a \to 0^+$ represents increasingly severe collapse (Haeffele et al., 2021). Here, the same contraction is applied to every view pair $v \neq r$, and the relations hold term-wise in the sum. Consider the operator:*

$$\mathbf{G}(a\mathbf{Z}^{vr}) = a^2\mathbf{G}(\mathbf{Z}^{vr}), \qquad D(a\mathbf{Z}^{vr}) = a^2 D(\mathbf{Z}^{vr}). \tag{7}$$

Thus, the operator and MSE term share the same quadratic factor, which cancels:

$$\mathcal{L}_{\text{RDE-MSE}}(a\mathbf{Z}^{vr}) = \frac{1}{N}\sum_{v\neq r} \frac{a^2\left\langle (\mathbf{Z}^{vr})^{\top}\mathbf{Z}^{vr}, (\mathbf{C}^{vr}-\mathbf{I})(\mathbf{C}^{vr}-\mathbf{I})^{\top}\right\rangle}{a^2 D(\mathbf{Z}^{vr})} = \mathcal{L}_{\text{RDE-MSE}}(\mathbf{Z}^{vr}). \tag{8}$$

A detailed proof is provided in Appendix A.2. Thus, the quadratic factor introduced by the ordinary MSE term is canceled by the identical factor in the dispersion operator. Consequently, optimizing Eq. (6) no longer favors shrinking $\mathbf{Z}^{vr}$, thereby mitigating MSE-alignment-induced representation collapse and allowing the alignment matrices to learn accurate sample-wise matching degrees across views for more effective cross-view alignment. The aligned representations $\mathbf{P}^{vr}\mathbf{Z}^v$ and $\mathbf{Z}^r$ are then used for consistency fusion.

3.4 Representation Dispersion-Enhanced Contrastive Learning (RDE-CL)

To further exploit cross-view consistency, we design RDE-CL to dynamically adapt the contrastive temperature according to the joint dispersion of each view pair, thereby promoting stable cross-view consistency. First, following conventional contrastive learning, we define cross-view positive and negative pairs from the aligned representations. Specifically, for each non-reference view v, we define $\widetilde{\mathbf{Z}}^v = \mathbf{P}^{vr}\mathbf{Z}^v$ as aligned representations. For each sample index i, $(\widetilde{\mathbf{z}}_i^v, \mathbf{z}_i^r)$ forms a positive pair, while $\{(\widetilde{\mathbf{z}}_i^v, \mathbf{z}_j^r) \mid j \neq i\}$ forms the corresponding negative pairs. Positive pairs encourage high similarity between matched samples, whereas negative pairs suppress similarity between mismatched samples, thereby promoting cross-view consistency and inter-sample discrimination.

However, a fixed contrastive temperature applies the same logit scaling throughout optimization, regardless of changes in the joint dispersion of each view pair, which prevents contrastive learning from adaptively adjusting the contrastive temperature according to the joint dispersion. Based on this limitation, we set the effective temperature as

$$\tau_{\text{eff}}^{v} = \tau_{\text{base}} D(\mathbf{Z}^{vr}), \tag{9}$$

where $\tau_{\text{base}} > 0$ is a fixed base temperature, and τ_{eff}^{v} is recomputed as the representations are updated. When joint dispersion decreases, indicating a more degenerate joint representation, Eq. (9) lowers the temperature, thereby amplifying existing differences between positive and negative cosine similarities and strengthening the contrastive signal. Using this state-dependent temperature, the RDE-CL objective is

$$\mathcal{L}_{\text{RDE-CL}} = -\frac{1}{N} \sum_{v \neq r} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}\left(\widetilde{\mathbf{z}}_i^v, \mathbf{z}_i^r\right) / \tau_{\text{eff}}^v)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(\widetilde{\mathbf{z}}_i^v, \mathbf{z}_j^r\right) / \tau_{\text{eff}}^v\right)}, \tag{10}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. By adjusting the logit scale according to the current joint dispersion, RDE-CL maintains a balanced contrastive signal as the representations evolve and promotes more reliable cross-view consistency.

Finally, we fuse the aligned representations as $\mathbf{Z}_{\text{fused}} = \frac{1}{V}\left(\mathbf{Z}^r + \sum_{v \neq r} \widetilde{\mathbf{Z}}^v\right)$, and apply K-means. The complete loss of RDE-UMC is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{A}} \mathcal{L}_{\text{RDE-MSE}} + \lambda_{\text{C}} \mathcal{L}_{\text{RDE-CL}}, \tag{11}$$

where λ_{A} and λ_{C} balance cross-view alignment and consistency learning, respectively.

Algorithm 1 The Training Process of RDE-UMC.

Input: Unaligned data $\mathcal{X} = \{\mathbf{X}^v\}_{v=1}^V$, clusters C, weights $\lambda_{\text{A}}, \lambda_{\text{C}}$, temperature τ_{base}, and epochs $T_{\text{pre}}, T_{\text{joint}}$.
Output: Clustering results $\mathbf{y}$.
Process:
1: Initialize and pretrain the autoencoders for T_{pre} epochs using Eq. (3) to obtain $\{\mathbf{Z}^v\}_{v=1}^V$.
2: Select the reference view $r \leftarrow \arg\max_v D_v$ and initialize the mapping matrices $\{\mathbf{P}^{vr}\}_{v \neq r}$.
3: **for** $t = 1, \ldots, T_{\text{joint}}$ **do**
4: Compute $D(\mathbf{Z}^{vr})$, RDE-MSE, and RDE-CL using Eqs. (5), (6), and (10).
5: Jointly update the autoencoders and mapping matrices using the total objective in Eq. (11).
6: **end for**
7: Fuse the aligned representations and perform K-means to obtain $\mathbf{y}$.

3.5 Optimization

Algorithm 1 summarizes the training process. After pretraining the view-specific autoencoders, selecting the reference view, and initializing the mapping matrices, we jointly update the autoencoders and mapping matrices using the objective in Eq. (11), while computing the dispersion operator, RDE-MSE, and RDE-CL at each iteration. Finally, we fuse the aligned representations and apply K-means. During RDE-CL computation, $D(\mathbf{Z}^{vr})$ is treated as a stop-gradient scalar when forming τ_{eff}^{v}, whereas it remains differentiable in RDE-MSE.

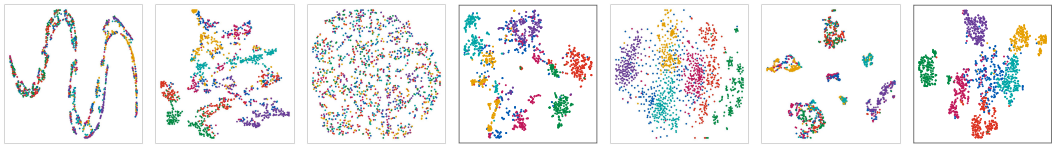
4 Experiments

4.1 Experimental Settings

Datasets. We evaluate RDE-UMC on eight widely used multi-view benchmark datasets **(Table 1)**: **MSRCv1** (Winn & Jojic, 2005), **BBCSport** (Greene & Cunningham, 2006), **REU** (Amini et al., 2009), **Caltech-5V-7** (Fei-Fei et al., 2004), **Handwritten** (van Breukelen et al., 1998), **Scene15** (Fei-Fei & Perona, 2005), **Hdigit** (Chen et al., 2022), and **Aloi-100** (Geusebroek et al., 2005).

Table 1: Statistics of the benchmark datasets.

Data	Samples	Clusters	View dimensions
MSRCv1	210	7	24/576/512/256/254
BBCSport	544	5	3183/3224
REU	1,200	6	2000/2000/2000/2000/2000
Caltech-5V-7	1,400	7	40/254/1984/512/928
Handwritten	2,000	10	240/76/216/47/64/6
Scene15	4,485	15	20/59/40
Hdigit	10,000	10	784/256
Aloi-100	10,800	100	77/13/64/125



(a) MVC-UM (b) UPMGC (c) MGCCFF (d) JHCL (e) GLANCE (f) DAGF (g) Ours

Figure 3: t-SNE visualizations of the clustering results for different methods on Caltech-5V-7.

Compared methods. We compare RDE-UMC with six recent UMC methods: **MVC-UM** (Yu et al., 2021), **UPMGC** (Wen et al., 2024), **MGCCFF** (Zhao et al., 2025), **JHCL** (Jiang et al., 2025), **GLANCE** (Xin et al., 2026), and **DAGF** (Jiang et al., 2026).

Evaluation metrics. We evaluate clustering with four standard metrics: Accuracy (ACC), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Purity (PUR), whose detailed definitions are provided in (Liang et al., 2022). For all these metrics, higher values indicate better performance.

Implementation details. RDE-UMC is implemented in PyTorch (Paszke et al., 2019) and optimized with Adam (Kingma & Ba, 2015). Each view uses an encoder with dimensions $\{D_v, 500, 500, 2000, d\}$, a symmetric decoder, and ReLU hidden activations; the autoencoders are pretrained for 100 epochs and jointly optimized for 150 epochs. Experiments use two Intel Xeon Gold 6148 CPUs, eight NVIDIA GeForce RTX 3090 GPUs, and 512 GB memory.

4.2 Experimental Results

Table 2 compares the clustering results on eight datasets; bold and underlined values denote the best and second-best results, respectively. Figure 3 presents the corresponding t-SNE visualizations on Caltech-5V-7. Together, the table and figure support the following observations.

(1) In Table 2, RDE-UMC is superior to all compared methods on nearly all evaluation metrics across the eight datasets and has a clear leading gap over the second-best methods. Especially on the Caltech-5V-7 dataset, the improvements over the second-best method, GLANCE, are 14.50, 8.64, 14.35, and 9.21 percentage points on ACC, NMI, ARI, and PUR, respectively. These experimental results strongly demonstrate the effectiveness and superiority of RDE-UMC, and we attribute such success to our designed covariance-volume-based dispersion enhancement operator, which mitigates MSE-alignment-induced representation collapse through dispersion-normalized alignment and adaptive temperature, preserving discriminative representations for consistency fusion and clustering.

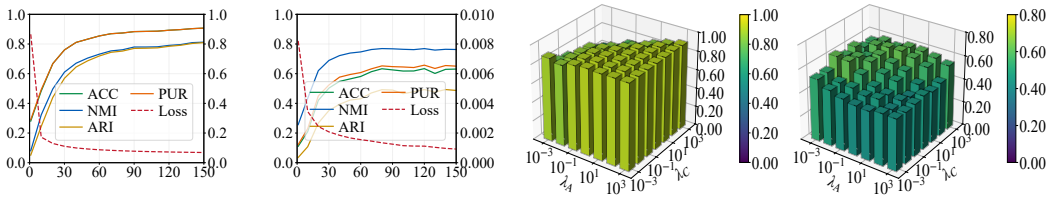
(2) In Figure 3, we select six representative unaligned multi-view clustering methods to conduct t-SNE visualization comparisons with our proposed RDE-UMC. According to the visualization results, the compared baseline methods exhibit varying degrees of mixed or diffuse cluster distributions, whereas RDE-UMC exhibits a clearer and more cohesive cluster structure. This demonstrates the superiority of RDE-UMC in preserving discriminative aligned representations and fusing cross-view information from unaligned multi-view data.

4.3 Model Analysis

Parameter sensitivity analysis. We experimentally evaluate the effects of λ_{A} and λ_{C} on the clustering performance of RDE-UMC. Figures 4(c) and (d) show the ACC values on Caltech-5V-7

Table 2: Clustering results on eight fully unaligned multi-view datasets. The best result is in bold and the second-best result is underlined.

Dataset	Metric	Method						
		MVC-UM	UPMGC	MGCCFF	JHCL	GLANCE	DAGF	Ours
MSRCv1	ACC	0.3381	0.5429	0.4069	0.5333	0.1952	0.3286	**0.7810**
	NMI	0.2290	0.4742	0.2964	0.5332	0.0630	0.2325	**0.6243**
	ARI	0.1112	0.3452	0.1678	0.3237	0.0000	0.2392	**0.5630**
	PUR	0.3714	0.5762	0.4174	0.5476	0.1952	0.3619	**0.7810**
BBCSport	ACC	0.4100	0.4492	0.3971	0.3647	0.3985	0.3566	**0.8051**
	NMI	0.1800	0.2948	0.1356	0.0325	0.0884	0.0154	**0.5851**
	ARI	0.0698	0.1933	0.0729	0.0015	0.0309	0.0011	**0.6424**
	PUR	0.4233	0.5214	0.4055	0.3721	0.4096	0.3603	**0.8051**
REU	ACC	0.4083	0.5440	0.3145	0.1698	0.2917	0.1707	**0.6333**
	NMI	0.2373	0.3149	0.1137	0.0084	0.1338	0.0089	**0.3993**
	ARI	0.1756	0.2682	0.0777	0.0000	0.0600	0.0000	**0.3424**
	PUR	0.4619	0.5588	0.3258	0.1710	0.3000	0.1715	**0.6333**
Caltech-5V-7	ACC	0.3200	0.5049	0.2057	0.6550	0.7621	0.4907	**0.9071**
	NMI	0.1633	0.4129	0.0347	0.5113	0.7275	0.4045	**0.8139**
	ARI	0.0979	0.3200	0.0174	0.4312	0.6670	0.2809	**0.8105**
	PUR	0.3393	0.5526	0.2236	0.6643	0.8150	0.5129	**0.9071**
Handwritten	ACC	0.7465	0.6404	0.7248	0.6410	0.1606	0.6130	**0.8545**
	NMI	0.7230	0.6485	**0.7543**	0.6531	0.1013	0.6422	0.7530
	ARI	0.6305	0.5043	0.5622	0.5051	0.0211	0.4080	**0.7139**
	PUR	0.7465	0.6652	0.7248	0.6410	0.1616	0.6290	**0.8545**
Scene15	ACC	0.2608	0.1386	0.1529	0.3304	0.3096	0.1608	**0.3960**
	NMI	0.2787	0.0576	0.1254	0.3654	0.3013	0.1501	**0.3673**
	ARI	0.1121	0.0434	0.0224	0.1795	0.1455	0.1337	**0.2097**
	PUR	0.2874	0.1503	0.1749	0.3320	0.3526	0.1663	**0.4234**
Hdigit	ACC	0.4627	0.4487	0.4167	0.2673	0.1842	0.3577	**0.7594**
	NMI	0.4418	0.4187	0.4690	0.2708	0.1144	0.4526	**0.6806**
	ARI	0.3987	0.2897	0.0626	0.0772	0.0201	0.2193	**0.6176**
	PUR	0.5031	0.4887	0.5182	0.2779	0.1888	0.3585	**0.7607**
Aloi-100	ACC	0.3543	0.0383	0.4176	0.0527	0.4806	0.2495	**0.6311**
	NMI	0.6039	0.1118	0.6877	0.1056	0.6802	0.5745	**0.7630**
	ARI	0.2381	0.0213	0.0747	0.0002	0.2730	0.1192	**0.4871**
	PUR	0.3692	0.0402	0.4252	0.0623	0.5159	0.2560	**0.6508**



(a) Caltech-5V-7 (b) Aloi-100 (c) Caltech-5V-7 (d) Aloi-100

Figure 4: Convergence analysis in (a) and (b), and parameter sensitivity analysis in (c) and (d).

and Aloi-100 as both parameters vary from 10^{-3} to 10^{3}. It can be observed from the results that RDE-UMC is relatively insensitive to both parameters when λ_{C} is one order of magnitude larger than λ_{A}, with λ_{A} varying from 10^{-2} to 10^{2} and λ_{C} varying from 10^{-1} to 10^{3} on both datasets.

Convergence analysis. Figures 4(a) and (b) show the convergence curves of RDE-UMC on Caltech-5V-7 and Aloi-100, including the training loss and four evaluation metrics in each subfigure. The loss drops rapidly and then stabilizes, while the metrics fluctuate early before progressively stabilizing, indicating stable training on both datasets.

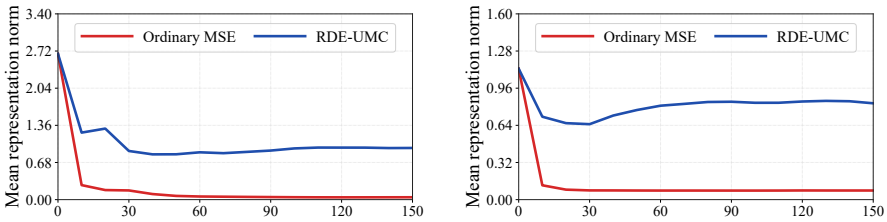
Ablation studies. We conduct two series of ablation studies from the perspectives of loss functions and the representation dispersion enhancement operator on Caltech-5V-7 and Aloi-100. Table 3 presents the loss-function ablation, where $\mathcal{L}_{\mathrm{rec}}$ reconstructs the input views, $\mathcal{L}_{\mathrm{RDE\text{-}MSE}}$ performs

Table 3: Loss-function ablation on Caltech-5V-7 and Aloi-100.

	Loss			Caltech-5V-7				Aloi-100			
	$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{RDE-MSE}}$	$\mathcal{L}_{\text{RDE-CL}}$	**ACC**	**NMI**	**ARI**	**PUR**	**ACC**	**NMI**	**ARI**	**PUR**
(A)	✓		✓	0.4764	0.3223	0.2416	0.4764	0.2794	0.4988	0.1561	0.3016
(B)	✓	✓		0.8971	0.7969	0.7916	0.8971	0.3069	0.5090	0.1681	0.3413
(C)	✓	✓	✓	**0.9071**	**0.8139**	**0.8105**	**0.9071**	**0.6311**	**0.7630**	**0.4871**	**0.6508**

Table 4: Operator ablation on Caltech-5V-7 and Aloi-100.

	Dispersion operator		Caltech-5V-7				Aloi-100			
	RDE-MSE	**RDE-CL**	**ACC**	**NMI**	**ARI**	**PUR**	**ACC**	**NMI**	**ARI**	**PUR**
(a)			0.8329	0.6839	0.6643	0.8329	0.4883	0.6783	0.3448	0.5125
(b)		✓	0.8607	0.7455	0.7094	0.8607	0.5227	0.6903	0.3743	0.5584
(c)	✓		0.8957	0.7935	0.7888	0.8957	0.5673	0.7461	0.4271	0.5910
(d)	✓	✓	**0.9071**	**0.8139**	**0.8105**	**0.9071**	**0.6311**	**0.7630**	**0.4871**	**0.6508**



(a) Caltech-5V-7 (b) Aloi-100

Figure 5: Mean representation-norm trajectories during training.

dispersion-normalized cross-view alignment, and $\mathcal{L}_{\text{RDE-CL}}$ performs dispersion-adaptive contrastive learning. Table 4 further reports the use of the same dispersion enhancement operator in the two branches: the RDE-MSE column indicates whether it is used for dispersion normalization, and the RDE-CL column indicates whether it is used for temperature adaptation. According to Tables 3 and 4, we make the following observations.

(1) In Table 3, variant (C) outperforms (A) on both datasets, showing the benefit of adding RDE-MSE to RDE-CL for cross-view alignment. It also surpasses (B), demonstrating that RDE-CL provides complementary gains beyond RDE-MSE by exploiting cross-view consistency.

(2) In Table 4, variant (d) outperforms (a), confirming the overall benefit of incorporating the operator in both branches. Its improvement over (b) shows the contribution of dispersion normalization in RDE-MSE, while its improvement over (c) shows the contribution of adaptive temperature in RDE-CL. These results verify that the two uses of the same operator provide complementary benefits for cross-view alignment and consistency learning.

Representation norm analysis. To assess representation collapse caused by ordinary MSE alignment, we compare RDE-UMC with an ordinary MSE control on Caltech-5V-7 and Aloi-100, while keeping all other settings fixed. As shown in Figure 5, the ordinary MSE control rapidly causes the representation norm to approach zero, whereas RDE-UMC maintains a larger representation scale. These results show that RDE-UMC mitigates MSE-alignment-induced collapse while preserving separable clusters for alignment and consistency fusion.

5 Conclusion

In this paper, we propose RDE-UMC to mitigate MSE-alignment-induced representation collapse in unaligned multi-view clustering. The method introduces a representation dispersion enhancement operator that captures degeneration across principal directions while preserving representation separability. Incorporated into ordinary MSE alignment, the operator yields RDE-MSE for reliable cross-view alignment; in RDE-CL, it adapts the contrastive temperature to promote consistency fusion. Theoretical analysis, extensive comparisons with state-of-the-art methods, and further analyses on benchmark datasets verify the effectiveness and superiority of RDE-UMC.

AI USE STATEMENT

Generative AI tools were used solely to assist with writing and language polishing of this manuscript. All AI-assisted text was carefully reviewed and revised by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Massih R. Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views: An application to multilingual text categorization. In *Advances in Neural Information Processing Systems*, pp. 28–36, 2009.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

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

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 3015–3024, 2021.

Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 12350–12368, 2023.

Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 524–531, 2005.

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 *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 178–178, 2004.

Hongchang Gao, Feiping Nie, Xuelong Li, and Heng Huang. Multi-view subspace clustering. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4238–4246, 2015.

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

Derek Greene and Pádraig Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *International Conference on Machine Learning*, pp. 377–384, 2006.

Benjamin D. Haeffele, Chong You, and Rene Vidal. A critique of self-expressive deep subspace clustering. In *International Conference on Learning Representations*, pp. 1–20, 2021.

Jintian Ji, Songhe Feng, and Yidong Li. Tensorized unaligned multi-view clustering with multi-scale representation learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1246–1256, 2024.

Hongyu Jiang, Yuxin Huo, Sirou Sheng, Hong Tao, and Chenping Hou. Scalable one-step unaligned multi-view clustering via joint high-order correlation learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1414–1422, 2025.

Hongyu Jiang, Hong Tao, Zhangqi Jiang, and Chenping Hou. Unaligned multi-view clustering via diversified anchor graph fusion. *Pattern Recognition*, 170:111977, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015.

Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems*, pp. 1413–1421, 2011.

Youwei Liang, Dong Huang, Chang Dong Wang, and S Yu Philip. Multi-view graph learning by joint modeling of consistency and inconsistency. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2):2848–2862, 2022.

Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pp. 252–260, 2013.

Xinwang Liu, Xinzhong Zhu, Miaomiao Li, Lei Wang, Chang Tang, Jianping Yin, Dinggang Shen, Huaimin Wang, and Wen Gao. Late fusion incomplete multi-view clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2410–2423, 2019.

Serdar Ozsoy, Shadi Hamdan, Sercan Ö. Arik, Deniz Yuret, and Alper T. Erdogan. Self-supervised learning with an information maximization criterion. In *Advances in Neural Information Processing Systems*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Daniel J. Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1255–1265, 2021.

Martijn van Breukelen, Robert P. W. Duin, David M. J. Tax, and J. E. den Hartog. Handwritten digit recognition by combined classifiers. *Kybernetika*, 34(4):381–386, 1998.

Hao Wang, Yan Yang, and Bing Liu. GMC: Graph-based multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1116–1129, 2020.

Jing Wang and Songhe Feng. Contrastive and view-interaction structure learning for multi-view clustering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5055–5063, 2024.

Yi Wen, Siwei Wang, Qing Liao, Weixuan Liang, Ke Liang, Xinhang Wan, and Xinwang Liu. Unpaired multi-view graph clustering with cross-view structure matching. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16049–16063, 2024.

John M. Winn and Nebojsa Jojic. LOCUS: Learning object classes with unsupervised segmentation. In *IEEE International Conference on Computer Vision*, pp. 756–763, 2005.

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.

Hong Yu, Jia Tang, Guoyin Wang, and Xinbo Gao. A novel multi-view clustering method for unknown mapping relationships between cross-view samples. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2075–2083, 2021.

Lianjin Yu, Bohang Sun, Xiangning Zeng, Haobo Wang, and Gengyu Lyu. Hypergraph-based unaligned multi-view clustering via cluster-aware feature extraction. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6282–6291, 2026.

Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2921–2927, 2017.

Liang Zhao and Qiongjie Xie. Distribution-level multi-view clustering for unaligned data. *IEEE Signal Processing Letters*, 31:2330–2334, 2024.

Liang Zhao, Ziyue Wang, Xiao Wang, Zhikui Chen, and Bo Xu. Incomplete and unpaired multi-view graph clustering with cross-view feature fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 22786–22794, 2025.

Linlin Zong, Xianchao Zhang, Xinyue Liu, and Hong Yu. Multi-view clustering on data with partial instances and clusters. *Neural Networks*, 129:19–30, 2020.

A Proofs and Additional Analysis

A.1 MSE-Alignment-Induced Representation Collapse

Proof of Theorem 1. By the definitions of $\mathbf{Z}^{vr}$ and $\mathbf{C}^{vr}$, direct block-matrix multiplication gives

$$\mathbf{Z}^{vr}\mathbf{C}^{vr} = \left[(\mathbf{Z}^v)^\top \quad (\mathbf{P}^{vr}\mathbf{Z}^v)^\top\right]. \tag{12}$$

Subtracting $\mathbf{Z}^{vr} = [(\mathbf{Z}^v)^\top, (\mathbf{Z}^r)^\top]$ from both blocks yields

$$\mathbf{Z}^{vr}\mathbf{C}^{vr} - \mathbf{Z}^{vr} = \left[\mathbf{0} \quad (\mathbf{P}^{vr}\mathbf{Z}^v - \mathbf{Z}^r)^\top\right]. \tag{13}$$

The Frobenius norm is invariant to transposition and to appending a zero block, so

$$\|\mathbf{P}^{vr}\mathbf{Z}^v - \mathbf{Z}^r\|_F^2 = \|\mathbf{Z}^{vr}\mathbf{C}^{vr} - \mathbf{Z}^{vr}\|_F^2 = \left\langle (\mathbf{Z}^{vr})^\top\mathbf{Z}^{vr}, (\mathbf{C}^{vr} - \mathbf{I})(\mathbf{C}^{vr} - \mathbf{I})^\top \right\rangle. \tag{14}$$

Summing over $v \neq r$ and dividing by N gives Eq. (2). For any $a \in (0, 1)$, let $\widetilde{\mathbf{Z}}^{vr} = a\mathbf{Z}^{vr}$. The ordinary MSE alignment loss then satisfies

$$\begin{aligned}\mathcal{L}_{\mathrm{MSE}}(\widetilde{\mathbf{Z}}^{vr}) &= \frac{1}{N}\sum_{v\neq r}\left\langle (a\mathbf{Z}^{vr})^\top(a\mathbf{Z}^{vr}), (\mathbf{C}^{vr} - \mathbf{I})(\mathbf{C}^{vr} - \mathbf{I})^\top \right\rangle \\ &= a^2\mathcal{L}_{\mathrm{MSE}}(\mathbf{Z}^{vr}).\end{aligned} \tag{15}$$

Hence, the ordinary MSE alignment loss creates a scale-shrinking shortcut during optimization: minimizing it can be achieved by decreasing a and jointly shrinking $\mathbf{Z}^{vr}$. As $a \to 0^+$, the optimization causes $\mathbf{Z}^{vr}$ to approach zero, where every alignment term vanishes. Thus, optimizing the ordinary MSE alignment loss can induce representation collapse.

Implication for alignment and clustering. MSE-alignment-induced representation collapse compresses the pairwise distances $\|\mathbf{z}_i^v - \mathbf{z}_j^v\|_2$ as $\mathbf{Z}^v$ approaches zero, although the resulting zero alignment loss does not indicate recovery of the unknown cross-view correspondences. The representations then provide no reliable geometric evidence for distinguishing matched from unmatched samples. Moreover, the simultaneous contraction of within-cluster and between-cluster distances distorts the cluster structure required by the subsequent clustering stage.

Effect of the reconstruction loss. The joint objective also contains the reconstruction loss, so we further examine whether reconstruction prevents the scale-shrinking shortcut identified above. Reconstruction requires the latent representation to retain sufficient sample-distinguishing information for recovering each input. This requirement prevents distinct inputs from sharing an exactly identical code when they must be reconstructed differently, but it does not impose a positive lower bound on the distances between their codes or on the overall latent scale. For a finite dataset, a sufficiently expressive encoder–decoder network can therefore place distinct samples arbitrarily close in the latent space and still decode them accurately, provided that their latent representations remain distinguishable. Haeffele et al. (2021) show that even simple encoder–decoder networks can produce arbitrarily close latent samples while retaining exact reconstruction under mild architectural conditions.

The scale reparameterization of the encoder–decoder makes this freedom explicit. The architecture used here ends each encoder with a linear layer and begins each decoder with an affine layer. Let $\mathbf{W}_D^v$ and $\mathbf{b}_D^v$ denote the weight and bias of the first decoder layer. For any $a \in (0, 1)$, jointly scale the latent representation and compensate the decoder weight by defining

$$\widetilde{\mathbf{Z}}^v = a\mathbf{Z}^v, \qquad \widetilde{\mathbf{W}}_D^v = a^{-1}\mathbf{W}_D^v.$$

The input to the first decoder activation then remains unchanged:

$$\widetilde{\mathbf{Z}}^v(\widetilde{\mathbf{W}}_D^v)^\top + \mathbf{b}_D^v = (a\mathbf{Z}^v)(a^{-1}\mathbf{W}_D^v)^\top + \mathbf{b}_D^v = \mathbf{Z}^v(\mathbf{W}_D^v)^\top + \mathbf{b}_D^v. \tag{16}$$

Thus, every subsequent decoder layer receives the same input as before, so the reconstructed samples and the reconstruction loss are unchanged. Taking $a \to 0^+$ contracts the latent representations while the inverse change in the first decoder weight preserves reconstruction.

By contrast, Theorem 1 shows that optimizing the ordinary MSE alignment term along this scale-reparameterization direction can cause the representations to collapse toward zero, thereby reducing the alignment term to zero. Consequently, for $\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda\mathcal{L}_{\text{MSE}}$, we have

$$\mathcal{L}_{\text{rec}} \text{ remains unchanged}, \qquad \mathcal{L}_{\text{MSE}} \to 0, \qquad \mathcal{L} \to \mathcal{L}_{\text{rec}}. \tag{17}$$

At the exactly collapsed point, identical latent codes cannot reconstruct distinct inputs. The relevant failure mode is therefore asymptotic: the codes remain distinguishable but become arbitrarily close, while the decoder weights compensate for their decreasing scale. Thus, reconstruction can preserve input recovery without preventing MSE-alignment-induced representation collapse.

A.2 Quadratic Scaling of the Representation Dispersion Operator

Proof of Theorem 2. Let $\widetilde{\mathbf{Z}}^{vr} = a\mathbf{Z}^{vr}$ with $0 < a < 1$. Substituting this scaled pair into Eq. (4) gives

$$\begin{aligned}
\mathbf{G}(\widetilde{\mathbf{Z}}^{vr}) &= \frac{1}{2N}\widetilde{\mathbf{Z}}^{vr}\mathbf{H}_{2N}(\widetilde{\mathbf{Z}}^{vr})^\top \\
&= \frac{1}{2N}(a\mathbf{Z}^{vr})\mathbf{H}_{2N}(a\mathbf{Z}^{vr})^\top \\
&= a^2\frac{1}{2N}\mathbf{Z}^{vr}\mathbf{H}_{2N}(\mathbf{Z}^{vr})^\top = a^2\mathbf{G}(\mathbf{Z}^{vr}).
\end{aligned}$$

Thus, every entry of the joint covariance matrix is multiplied by a^2. For the $d \times d$ covariance matrix, the determinant is consequently

$$\det(\mathbf{G}(a\mathbf{Z}^{vr})) = \det\big(a^2\mathbf{G}(\mathbf{Z}^{vr})\big) = (a^2)^d \det(\mathbf{G}(\mathbf{Z}^{vr})) = a^{2d}\det(\mathbf{G}(\mathbf{Z}^{vr}))\,.$$

Applying the d-th root in the definition of the dispersion operator therefore gives

$$\begin{aligned}
D(a\mathbf{Z}^{vr}) &= \det(\mathbf{G}(a\mathbf{Z}^{vr}))^{1/d} \\
&= \big(a^{2d}\det(\mathbf{G}(\mathbf{Z}^{vr}))\big)^{1/d} \\
&= a^2\det(\mathbf{G}(\mathbf{Z}^{vr}))^{1/d} = a^2 D(\mathbf{Z}^{vr}).
\end{aligned}$$

Equivalently, the eigenvalues of $\mathbf{G}(a\mathbf{Z}^{vr})$ are $a^2\lambda_k^{vr}$, so their geometric mean is also multiplied by a^2. Hence, D is homogeneous of degree two in the representation pair, matching the quadratic scaling of the ordinary MSE alignment numerator. Specifically, the scaled RDE-MSE objective satisfies

$$\begin{aligned}
\mathcal{L}_{\text{RDE-MSE}}(a\mathbf{Z}^{vr}) &= \frac{1}{N}\sum_{v\neq r}\frac{\big\langle (a\mathbf{Z}^{vr})^\top(a\mathbf{Z}^{vr}), (\mathbf{C}^{vr}-\mathbf{I})(\mathbf{C}^{vr}-\mathbf{I})^\top\big\rangle}{D(a\mathbf{Z}^{vr})} \\
&= \frac{1}{N}\sum_{v\neq r}\frac{a^2\big\langle (\mathbf{Z}^{vr})^\top\mathbf{Z}^{vr}, (\mathbf{C}^{vr}-\mathbf{I})(\mathbf{C}^{vr}-\mathbf{I})^\top\big\rangle}{a^2 D(\mathbf{Z}^{vr})} \\
&= \mathcal{L}_{\text{RDE-MSE}}(\mathbf{Z}^{vr}).
\end{aligned}$$

Thus, the quadratic factor introduced by the MSE term is exactly canceled by the identical factor in the dispersion operator.

A.3 Covariance-Based Anti-Collapse Methods

Existing studies have explored covariance information to mitigate representation collapse induced by MSE-based representation matching. For example, VICReg retains the MSE-based invariance objective and introduces additional variance and covariance regularization to maintain non-degenerate representations (Bardes et al., 2022), while W-MSE transforms the learned representations using their covariance so that the representation dimensions are decorrelated and have unit variance before MSE matching (Ermolov et al., 2021). These methods mainly address collapse in the conventional MSE matching objective $\|\mathbf{z}_1 - \mathbf{z}_2\|_2^2$ by introducing additional regularization terms or transforming the learned representations. Instead, we study representation collapse induced by the mapping-based objective $\|\mathbf{P}^{vr}\mathbf{Z}^v - \mathbf{Z}^r\|_F^2$, where the cross-view mapping matrix and latent representations are jointly optimized, and incorporate the proposed representation dispersion enhancement operator directly into this objective, thereby mitigating collapse by counteracting its underlying mechanism.