Individuals Matter: Improving Deep Multi-View Clustering via Explicit Single-View Enhancement

Hanyang Li[1], **Yanzheng Wang**[1], **Yaxin Hou**[1], **Zhengxing Jiao**[1], **Hao Chen**[1], **Yuheng Jia**[1,2*]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
{lihanyang, 230260242, yaxin, jiaozx, haochen303, yhjia}@seu.edu.cn

Abstract

Although existing deep multi-view clustering methods often achieve high performance, they predominantly focus on improving the fused representation while overlooking the enhancement of individual views, often resulting in stagnant or inconsistently improved view-specific representations that limit the final fusion quality. To address this, we propose Explicit single-View Enhancement (EVE), a novel deep multi-view clustering framework designed to explicitly optimize the clustering capability of each individual view. Specifically, EVE employs an attention-based module to generate a global consensus as a reliable anchor, which subsequently guides the refinement of individual views through two key components: feature-level selective alignment and structural-level neighborhood propagation. Crucially, the enhanced single views will further boost the fused representation, and finally leading to a mutually reinforcing closed-loop feedback process. Our method is also naturally applicable to incomplete deep multi-view clustering scenarios. Extensive experiments on multiple benchmark datasets demonstrate that EVE consistently outperforms state-of-the-art competitors, achieving substantial average improvements of **5.9%** in ARI across all datasets. The code is available https://github.com/l-h-y168/EVE.

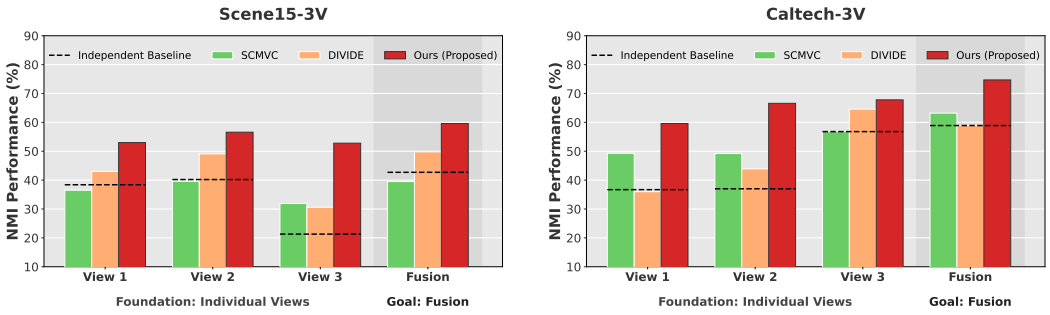
1 Introduction

Deep multi-view clustering studies how to cluster data by jointly learning representations from multiple views using deep neural networks in an unsupervised manner. By exploiting complementary information provided by different views, it enables more effective discovery of underlying cluster structures than using any single view alone [37, 1, 34]. Existing deep multi-view clustering methods can be roughly categorized into two paradigms based on their fusion mechanisms. Interaction-driven Fusion: Methods in this category [28, 11, 7] focus on establishing direct correspondences between views. They typically learn representations by maximizing the correlation or performing cross-view prediction between view pairs in a peer-to-peer manner. Consensus-driven Fusion: Methods in this category [30, 5, 8] explicitly construct a global consensus representation or a unified target during training. These methods adopt a "view-fuse" alignment strategy, constraining each individual view to align with this central anchor to ensure uniformity.

Despite their impressive performance in producing a high-quality fused representation, we find that behind this success lies a counter-intuitive phenomenon: while the fused representation becomes increasingly discriminative, the quality of individual view representations often stagnates or achieves inconsistent gains, failing to consistently benefit from the joint learning process. As empirically illustrated in Fig. 1, on Scene15-3V and Caltech-3V datasets, we compared the performance of views

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) (b)

Figure 1: **Motivation illustration.** We compare the NMI performance of individual view representations (labeled "Foundation") and the final fused representation (labeled "Goal") on (a) Scene15-3V and (b) Caltech-3V. The **black dashed lines** indicate the baseline performance: for individual views, they represent independent training; for the fusion group, they represent the performance of simple feature concatenation. **Key Observation:** While state-of-the-art methods like SCMVC (green) and DIVIDE (orange) achieve competitive fused performance, they bring inconsistent improvements to individual views, often hovering around or even falling below the independent baseline. In contrast, our method (red) explicitly enhances the performance of every single view, establishing a superior foundation that naturally leads to the best overall fusion performance.

trained independently (Single-View) against those trained within state-of-the-art multi-view frameworks SCMVC [30] and DIVIDE [28]. Surprisingly, we observe that these advanced frameworks yield **limited or inconsistent improvements** for individual views compared to the independent baseline. This observation suggests that current fusion mechanisms primarily function as a one-way aggregation, **failing to effectively leverage the high-quality global consensus to inversely rectify and optimize individual view representations**. We argue that a more effective multi-view system should close the loop: using the high-quality fused representation to explicitly correct and refine the individual views, which in turn boosts the final fusion upon this enhanced foundation.

To achieve the above goal, we propose Explicit single-View Enhancement (EVE), a novel deep multi-view clustering framework designed to explicitly rectify and optimize view-specific representations. Our method leverages the fused representation as a reliable global reference to explicitly guide the refinement of individual views. Specifically, we first introduce an attention-based fusion module to aggregate informative cues into a high-quality global fused representation. This representation serves as a reliable anchor to align all views to a unified semantic space, providing a solid foundation for further refinement. Subsequently, it guides individual views through two key components: (1) **Feature-level Guidance:** A loss-driven selective alignment strategy dynamically allows weaker views to distill knowledge from the most reliable view identified in the fused space, progressively enhancing view-specific features; (2) **Structure-level Guidance:** A fused-guided neighborhood propagation mechanism utilizes the consistent global structure to rectify the affinity matrices of individual views. By enhancing single-view quality, our framework naturally yields a more discriminative fused representation, which in turn provides more accurate supervision, forming a closed-loop optimization process. As illustrated in Fig. 1, by establishing the feedback loop, our method significantly improves the clustering performance of individual views, which in turn leads to a more reliable fused representation and further propel the overall multi-view clustering performance to a new state-of-the-art level. Moreover, the flexible attention fusion and the high-quality foundation built by single-view enhancement enable our method to seamlessly generalize to incomplete-view inference. In summary, our contributions are as follows.

- We rethink deep multi-view clustering from a single-view enhancement perspective, showing that improving the clustering quality of individual views is a crucial yet underexplored factor for achieving high multi-view clustering performance.
- We propose EVE, a deep multi-view clustering framework that forms a closed-loop interaction between fused and view-specific representations. It constructs a reliable global representation via attention-based fusion to refine individual views through feature-level selective alignment and structure-level neighborhood propagation, a design that enables flexible inference even with incomplete views.

- Extensive experiments on different multi-view datasets demonstrate that EVE consistently improves the clustering performance of individual views as well as the fused representation, achieving substantial average improvements of 5.9% in ARI across all datasets, and exhibiting strong robustness under incomplete-view scenarios.

2 Related Work

Clustering has been extensively studied from both multi-view learning and deep representation learning perspectives. Traditional multi-view clustering methods exploit complementary information and structural relationships across views [16], while deep clustering methods learn clustering-oriented representations in an unsupervised manner [31, 12, 17, 25, 22, 19]. Building on these two research directions, **deep multi-view clustering** (DMVC) integrates deep representation learning with multi-view information fusion to discover the underlying cluster structure more effectively. Early DMVC approaches primarily adopted autoencoders, focusing on reconstructing inputs to extract latent features [37, 24]. With the rapid development of self-supervised learning [13, 2, 10, 3, 36], recent methods have evolved to employ more sophisticated mechanisms for representation learning. Depending on how they integrate information and enforce consistency, existing methods can be broadly discussed under two paradigms: Interaction-driven Fusion and Consensus-driven Fusion. Interaction-driven fusion focuses on establishing direct correspondences between views to capture shared information. A dominant technique here is cross-view contrastive learning, where methods like [28, 11] maximize the similarity between positive pairs from different views while pushing apart negative ones. Beyond contrastive objectives, other principles are also widely adopted to enforce peer-to-peer consistency. For instance, some approaches maximize the mutual information between view pairs [26, 6] or utilize correlation-based objectives [7] to align feature distributions. These methods typically optimize the agreement between view-specific encoders directly. In contrast, consensus-driven fusion prioritizes constructing a unified target or global representation to guide the learning of individual views. Approaches such as [1, 30, 5, 32] typically employ specific fusion modules (e.g., attention mechanisms or weighted aggregation) to generate a fused embedding, which is then used as the supervision target for contrastive alignment. Furthermore, generative methods [9] and information bottleneck frameworks [33, 27] also fall into this scope, as they aim to map multiple views into a common latent space or a shared bottleneck variable, effectively treating this latent consensus as the basis for subsequent clustering.

Motivation. Despite these diverse technical routes, most existing methods focus on improving fused representation and typically overlook the enhancement of individual views. Motivated by this, we propose a novel framework, establishing a closed-loop framework where the fused global context serves as a reliable teacher to guide the refinement of each single view. By explicitly improving the feature and structural clarity of view-specific representations, we improve the clustering performance of individual views as well as the fused representation.

3 Method

We first provide the preliminary problem definition (**Sec. 3.1**) and framework overview (**Sec. 3.2**), and then detail the construction of a reliable global consensus via the **Attention-based Fusion Module (Sec. 3.3)**. Crucially, this consensus serves as a reference to explicitly enhance individual views through two complementary pathways: **Loss-driven Selective Alignment (Sec. 3.4)** for feature-level distillation and **Fused-guided Neighborhood Propagation (Sec. 3.5)** for structural-level rectification. Finally, the overall optimization and inference are presented in **Sec. 3.6**.

3.1 Problem Formulation

Given an unlabeled multi-view dataset $\mathcal{X} = \{X^{(1)}, X^{(2)}, \ldots, X^{(V)}\}$, deep multi-view clustering aims to partition the N samples into C clusters by jointly exploiting information from multiple views, without access to ground-truth labels. Here, each data sample is described by V views. $X^{(v)} \in \mathbb{R}^{N \times D_v}$ denotes the feature matrix of the v-th view, where N is the total number of samples and D_v is the feature dimension of that view.

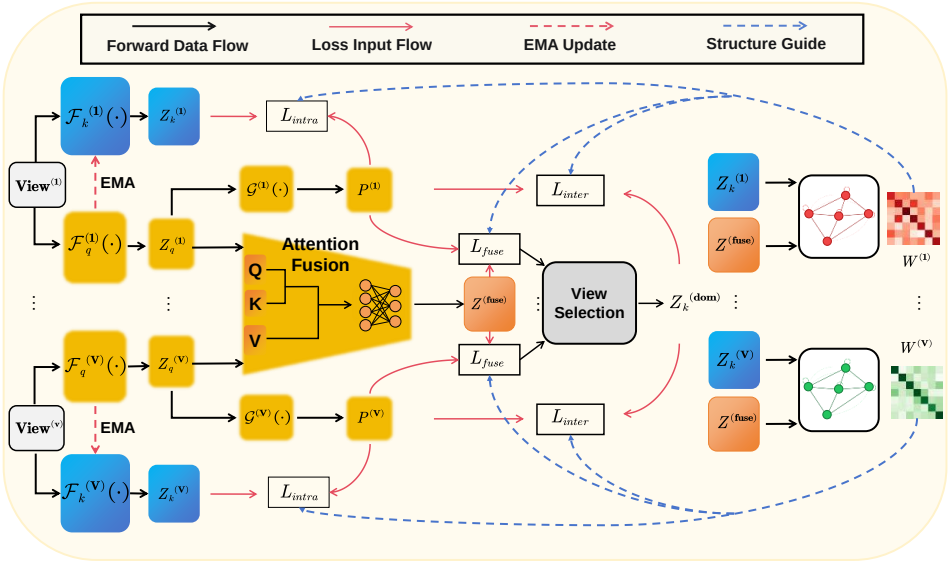


Figure 2: **Overview of EVE framework.** Features from online encoders are aggregated into a global $\mathbf{Z}^{(\text{fused})}$ via attention. This anchor refines individual views via: (1) **Feature-level Guidance**: *View Selection* identifies the dominant teacher for alignment ($\mathcal{L}_{\text{inter}}$); (2) **Structure-level Guidance**: *Structure Guide* (blue dashed lines) propagates global affinity to rectify targets for all objectives ($\mathcal{L}_{\text{intra}}$, $\mathcal{L}_{\text{fuse}}$, $\mathcal{L}_{\text{inter}}$). Finally, the clustering predictions are obtained by performing k-means on the learned global representation $\mathbf{Z}^{(\text{fused})}$.

3.2 Framework Overview

Network architecture. As shown in Fig. 2, for each view v, the model consists of an online encoder $\mathcal{F}_q^{(v)}(\cdot)$ and a projector $\mathcal{G}^{(v)}(\cdot)$, as well as a target encoder $\mathcal{F}_k^{(v)}(\cdot)$. The parameters of the target encoder are not directly optimized by backpropagation, but are updated as an exponential moving average (EMA) of the online encoder parameters. Given an input sample $x_i^{(v)} \in \mathbb{R}^{D_v}$ from the v-th view, the online branch produces a view-level feature and its projected feature, denoted as $\mathbf{z}_{q,i}^{(v)} \in \mathbb{R}^D$ and $\mathbf{p}_i^{(v)} \in \mathbb{R}^D$, respectively.In parallel, the target branch outputs the corresponding feature $\mathbf{z}_{k,i}^{(v)} \in \mathbb{R}^D$.

The view-specific feature vectors $\{\mathbf{z}_{q,i}^{(v)}\}_{v=1}^V$ from all online encoders are normalized and aggregated by an attention-based fusion module $\mathcal{A}(\cdot)$ to obtain a fused representation:

$$\mathbf{z}_i^{(\text{fused})} \in \mathbb{R}^D = \mathcal{A}\big(\mathbf{z}_{q,i}^{(1)}, \ldots, \mathbf{z}_{q,i}^{(V)}\big), \tag{1}$$

which serves as a global reference for subsequent cross-view alignment and representation enhancement.

Multi-level contrastive loss. For a batch of B samples, we adopt a general contrastive learning formulation that can operate between any two sets of L_2-normalized feature vectors $\{\mathbf{f}_{1,i}\}_{i=1}^B$ and $\{\mathbf{f}_{2,i}\}_{i=1}^B$ derived from the same batch. The contrastive loss for the i-th sample can be defined as:

$$\mathcal{L}_i = -\sum_{j=1}^{B} w_{ij} \log \frac{\exp\left(\langle \mathbf{f}_{1,i}, \mathbf{f}_{2,j} \rangle / \tau\right)}{\sum_{j'=1}^{B} \exp\left(\langle \mathbf{f}_{1,i}, \mathbf{f}_{2,j'} \rangle / \tau\right)}, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity between feature vectors, τ is a temperature parameter, w_{ij} denotes the weighting of the feature pair (i, j), and $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{B \times B}$ is the corresponding sample-wise weighting matrix for the batch. By default, $\mathbf{W}$ is initialized as the identity matrix $\mathbf{I}$, recovering the standard instance-level contrastive loss. Crucially, when the fused-guided neighborhood propagation (Sec. 3.5) is activated, we replace $\mathbf{I}$ with the refined affinity matrix $\mathbf{W}^{(v)}$ to inject structural guidance into all contrastive loss. The batch-wise loss is obtained by averaging over all

samples: $\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_i$. Based on different feature levels, we define three objectives, including an intra-view contrastive loss $\mathcal{L}_{\text{intra}}$, a fusion-guided contrastive loss $\mathcal{L}_{\text{fuse}}$, and an inter-view contrastive loss $\mathcal{L}_{\text{inter}}$. The three objectives are instantiated as follows.

The intra-view contrastive loss enforces consistency between the online and target features of the same view:

$$\mathcal{L}_{\text{intra}} = \frac{1}{V}\sum_{v=1}^{V}\mathcal{L}\big(\mathbf{P}^{(v)},\ \mathbf{Z}_k^{(v)}\big), \tag{3}$$

while the fusion-guided contrastive loss couples each view-level projected feature with the fused representation:

$$\mathcal{L}_{\text{fuse}} = \frac{1}{V}\sum_{v=1}^{V}\mathcal{L}\big(\mathbf{Z}^{(\text{fused})},\ \mathbf{P}^{(v)}\big). \tag{4}$$

In addition, the inter-view contrastive loss encourages cross-view alignment between different views:

$$\mathcal{L}_{\text{inter}} = \frac{1}{(V-1)}\sum_{v\neq u}\mathbf{M}^{(v,u)}\odot\mathcal{L}\big(\mathbf{P}^{(v)},\ \mathbf{Z}_k^{(u)}\big), \tag{5}$$

where $\mathbf{M}^{(v,u)} \in \{0,1\}^B$ is a binary selection mask vector that determines which samples participate in the alignment between view v and view u. The operation $\odot$ denotes the element-wise multiplication between the mask and the vector of sample-wise losses. The construction and refinement of $\mathbf{M}$ will be detailed in subsequent modules.

3.3 Attention-based Fusion Module

To fully exploit the complementary information across views, we introduce an attention-based fusion module. Unlike simple averaging or concatenation, this module allows each view to selectively attend to the informative features of other views, achieving sample-wise adaptive weighting. Formally, we stack the ℓ_2-normalized features of all views into a tensor and compute its query, key, and value projections denoted as $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{B\times V\times D}$ via linear transformations. To capture the complementary relationships across views, attention weights are computed along the view dimension:

$$\text{Attention}(\mathbf{Q},\mathbf{K}) = \text{softmax}\left(\mathbf{Q}\mathbf{K}^{\top}\right) \in \mathbb{R}^{B\times V\times V}, \tag{6}$$

The fused representation $\mathbf{z}_i^{(\text{fused})} \in \mathbb{R}^D$ for the i-th sample is then derived by aggregating the value features weighted by the attention score, followed by a multi-layer perceptron $\mathcal{M}(\cdot)$:

$$\mathbf{z}_i^{(\text{fused})} = \mathcal{M}\left(\sum_{v=1}^{V}\alpha_{i,v}\mathbf{v}_i^{(v)}\right), \tag{7}$$

where $\alpha_{i,v}$ denotes the learned importance (attention weight) of view v for sample i, and $\mathbf{v}_i^{(v)}$ represents the value embedding of view v. By selectively aggregating informative cues, this mechanism constructs a high-quality global reference, which provides the necessary guidance to refine individual views through loss-based selective alignment and fused-guided neighborhood propagation.

3.4 Loss-driven Selective Cross-view Alignment

While the fusion-guided loss $\mathcal{L}_{\text{fuse}}$ ensures global semantic consistency, direct cross-view alignment can further enhance discriminability. To avoid propagating noise from unreliable views, we leverage the fused representation as a reliability gauge to dynamically identify the most confident view (termed the **dominant view**), using it as a high-quality target to guide weaker ones. This strategy offers a natural optimization shortcut: since the dominant view serves as a high-quality, instance-level target within the same feature manifold, it is structurally easier for weaker views to capture. Furthermore, this direct alignment helps to mine and preserve distinct view-specific details that complement the aggregated global consensus. For each sample i, we first measure the relative reliability of each view v by computing a contrastive loss between the fused representation and the output of view-specific projector:

$$\ell_i^{(v)} = \mathcal{L}_i\big(\mathbf{z}_i^{(\text{fused})}, \mathbf{P}^{(v)}\big), \tag{8}$$

where $\mathbf{z}_i^{(\text{fuse})}$ is the fused feature from the attention-based module, and $\mathbf{P}^{(v)}$ is the projected feature of view v in the current batch. We then identify the dominant (strongest) view v_{dom} for each sample:

$$v_i^{\text{dom}} = \arg\min_v \ell_i^{(v)}. \tag{9}$$

A sample- and view-dependent mask $\mathbf{M}^{(v,u)}$ is constructed to allow selective alignment:

$$m_i^{(v,u)} = \begin{cases} 1, & \text{if } u = v_i^{\text{dom}} \text{ and } v \neq u, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

This ensures that each view v only aligns with the strongest view u for that sample i, and strong views are not forced to conform to weaker ones. The selective cross-view contrastive loss is then defined as:

$$\mathcal{L}_{\text{inter}} = \frac{1}{B(V-1)} \sum_{v \neq u} \sum_i m_i^{(v,u)} \mathcal{L}_i\big(\mathbf{p}_i^{(v)}, \mathbf{Z}_k^{(u)}\big). \tag{11}$$

It is worth noting that while the dominant view v^{dom} is not forced to align with other single views in $\mathcal{L}_{\text{inter}}$, it effectively serves as a teacher anchor. Crucially, v^{dom} is still optimized via $\mathcal{L}_{\text{fuse}}$ (Eq. 4). Since the fused representation aggregates complementary information from all views (including the weaker ones which may still contain unique structural cues), aligning v^{dom} with the fused embedding implicitly enhances it. This fosters a mutually reinforcing feedback loop: as weaker views are refined, they contribute to a cleaner fused representation. This enhanced global consensus, in turn, provides more precise supervision back to the dominant view, driving continuous improvement across all branches.

3.5 Fused-guided Neighborhood Propagation

To structurally refine individual views, we propose a fused-guided neighborhood propagation that computes the cross-space affinity between view-specific features and the fused representation. Given the features of the v-th view $\mathbf{Z}_k^{(v)} \in \mathbb{R}^{B \times D}$ and the fused representation $\mathbf{Z}^{(\text{fused})} \in \mathbb{R}^{B \times D}$, we first ℓ_2-normalize both features then compute the cross-view distance matrix between the target view and the fused representation as:

$$\mathbf{D}_{ij}^{(v)} = \left\| \mathbf{z}_{k,i}^{(v)} - \mathbf{z}_j^{(\text{fused})} \right\|_2^2 = 2 - 2 \left\langle \mathbf{z}_{k,i}^{(v)}, \mathbf{z}_j^{(\text{fused})} \right\rangle, \tag{12}$$

where $\mathbf{z}_{k,i}^{(v)}$ denotes the i-th sample in view v, and $\mathbf{z}_j^{(\text{fused})}$ denotes the j-th fused sample. The distance matrix is converted into an affinity matrix using an RBF-style kernel:

$$\mathbf{A}_{ij}^{(v)} = \exp\left(-\frac{\mathbf{D}_{ij}^{(v)}}{\sigma} \right), \tag{13}$$

where σ controls the kernel bandwidth. To ensure scale consistency and avoid trivial solutions, $\mathbf{A}^{(v)}$ is further normalized via Sinkhorn iterations to obtain a doubly stochastic affinity matrix $\tilde{\mathbf{W}}^{(v)}$. Finally, we perform fused-guided neighborhood propagation by combining the identity structure with multi-step affinity diffusion:

$$\mathbf{W}^{(v)} = \alpha \mathbf{I} + (1-\alpha) \left(\tilde{\mathbf{W}}^{(v)} \right)^{step}, \tag{14}$$

where $\mathbf{I}$ is the identity matrix, $\alpha \in [0,1]$ balances the original self-identity and the propagated neighborhood structure, and $step$ denotes the propagation depth. In all experiments, we set $\alpha = 0.5$ and $step = 5$. This design facilitates a progressive structural alignment: by quantifying the geometric discrepancy between the local view and the global consensus, the cross-space affinity directs the single view to gradually learn and adopt the more reliable global patterns. This effectively pulls the view-specific features towards the global semantic centers, correcting local distortions step-by-step during training.

Algorithm 1 The EVE algorithm

Require: Multi-view dataset $\mathcal{X} = \{\mathbf{x}^{(v)}\}_{v=1}^{V}$, batch size B.
Initialize: Online encoders $\mathcal{F}_q^{(v)}(\cdot)$, projectors $\mathcal{G}^{(v)}(\cdot)$, target encoders $\mathcal{F}_k^{(v)}(\cdot)$ for all views v, and fusion module $\mathcal{A}(\cdot)$.
while Clustering **do**
  for each batch $\{\mathbf{x}_i\}_{i=1}^{B}$ sampled from $\mathcal{X}$ **do**
    Extract representations $\{\mathbf{Z}_q^{(v)}, \mathbf{P}^{(v)}, \mathbf{Z}_k^{(v)}\}_{v=1}^{V}$ using online and target networks.
    Obtain the fused representation $\mathbf{Z}^{(\text{fused})}$ via attention-based fusion (Eq. 7).
    Construct the selective alignment mask $\mathbf{M}$ (Eq. 10).
    if fused-guided neighborhood propagation is activated **then**
      Compute the refined affinity matrix $\mathbf{W}^{(v)}$ (Eq. 14).
    else
      Set $\mathbf{W}^{(v)} = \mathbf{I}$.
    end if
    Compute the total training objective (Eq. 15)
    Update online encoders, projectors, and fusion modules via gradient descent, and update the target encoders via momentum.
  end for
end while
Perform k-means clustering on the fused representation $\mathbf{Z}^{(\text{fused})}$.

3.6 Optimization and Inference

Total objective. The overall optimization objective is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{intra}} + \mathcal{L}_{\text{fuse}} + \mathcal{L}_{\text{inter}}. \tag{15}$$

The specific training algorithm is presented in Algorithm 1, which integrates the attention-based fusion, selective alignment, and neighborhood propagation into a unified optimization framework.

Inference strategy. Our framework supports flexible inference. To handle incomplete-view scenarios, we leverage the attention mechanism by employing a masking strategy: attention weights for missing views are set to zero, allowing the model to adaptively aggregate only available views. Note that we utilize complete samples during training to ensure consensus reliability. Finally, regardless of view availability, the clustering predictions are obtained by performing standard k-means on the learned global representation $\mathbf{Z}^{(\text{fuse})}$.

4 Experiments

4.1 Experiment Settings

Datasets and baselines. We evaluated EVE on five widely used multi-view clustering datasets, including Scene15-2V, Scene15-3V [20], Caltech-3V [21], LandUse21 [35], and CCV [18], as summarized in the Appendix A, Table 4. Under the complete-view setting, we compare EVE with eight representative deep multi-view clustering approaches: DCP [26], GCFAgg [32], SCMVC [30], DIVIDE [28], CANDY [11], PLP [8], STCMC [14], COPER [7], MVP [9], and DFL-NET [4] providing a comprehensive evaluation against recent state-of-the-art methods. In addition, CANDY, DIVIDE, MVP, and PLP are also applicable to incomplete multi-view clustering, we further included them for comparison under the incomplete-view settings. More settings are shown in Appendix. A.

Evaluation metrics. We employed three commonly used clustering metrics to evaluate performance: Accuracy (ACC) [23], Normalized Mutual Information (NMI) [29], and Adjusted Rand Index (ARI) [15]. Higher scores indicate better clustering performance.

4.2 Main Results

Superior clustering performance on complete multi-view scenarios. Table 1 summarizes the quantitative results on five benchmark datasets. Our method EVE consistently achieves superior

Table 1: **Clustering performance (%) comparison on complete multi-view scenarios.** All results are reported as mean ± standard deviation over 5 runs with different random seeds. The best results are highlighted in **bold**, and the second-best results are in underlined.

Dataset	Metric	DCP (TPAMI 22)	GCFAgg (CVPR 23)	SCMVC (TMM 24)	DIVIDE (AAAI 24)	CANDY (NeurIPS 24)	PLP (TNNLS 24)	STCMC (AAAI 25)	COPER (ICLR 25)	MVP (ICLR 25)	DFL-NET (TKDE 25)	EVE (Ours)
Scene15-2V	ACC	39.9 ±1.7	40.5 ±0.5	40.6 ±0.9	50.2 ±0.8	43.8 ±0.9	40.9 ±1.1	37.5 ±1.1	41.2 ±1.3	45.3 ±0.6	36.1 ±1.9	**53.8** ±0.7
	NMI	42.3 ±1.8	40.0 ±0.3	41.8 ±0.3	48.7 ±0.4	42.7 ±0.5	44.2 ±0.3	42.9 ±1.2	43.6 ±0.7	43.8 ±0.5	40.7 ±1.9	**53.1** ±0.7
	ARI	22.2 ±2.7	22.5 ±0.6	24.2 ±0.5	32.7 ±0.7	26.7 ±0.5	26.7 ±0.3	23.1 ±1.0	24.8 ±0.6	26.6 ±0.4	21.7 ±1.4	**38.2** ±0.8
Scene15-3V	ACC	41.1 ±0.9	35.7 ±0.5	37.8 ±0.9	53.5 ±1.3	45.8 ±1.2	40.4 ±1.0	37.2 ±3.2	41.7 ±2.1	49.0 ±0.5	31.5 ±1.0	**60.7** ±0.8
	NMI	44.4 ±0.4	36.3 ±0.6	39.3 ±0.7	50.7 ±0.6	44.1 ±1.1	44.7 ±0.5	40.3 ±2.3	39.3 ±1.4	45.7 ±0.5	31.3 ±0.5	**58.3** ±0.9
	ARI	24.7 ±0.9	19.8 ±0.6	23.1 ±0.4	36.7 ±1.3	28.2 ±1.3	27.6 ±0.5	22.2 ±2.2	24.2 ±1.7	30.5 ±0.6	16.3 ±0.6	**45.1** ±0.9
Caltech-3V	ACC	53.8 ±2.4	68.9 ±1.4	73.5 ±0.9	75.2 ±1.8	72.9 ±3.1	73.0 ±0.7	63.7 ±6.4	64.4 ±4.0	76.9 ±1.3	72.7 ±3.8	**83.5** ±2.6
	NMI	51.7 ±1.4	56.8 ±2.1	63.4 ±1.1	62.9 ±2.3	58.0 ±2.6	65.4 ±0.5	58.0 ±4.5	56.1 ±3.7	61.4 ±2.2	65.9 ±3.5	**73.8** ±1.6
	ARI	37.9 ±2.0	49.8 ±1.7	55.0 ±1.0	58.6 ±2.5	53.7 ±3.6	57.1 ±0.6	45.1 ±9.2	47.3 ±4.3	57.3 ±2.2	59.0 ±5.2	**70.7** ±2.7
LandUse21	ACC	27.6 ±0.8	22.7 ±1.0	24.7 ±0.6	33.0 ±0.3	32.0 ±0.5	30.1 ±0.3	22.3 ±0.7	20.1 ±0.8	31.4 ±0.2	22.2 ±0.6	**34.4** ±0.7
	NMI	32.1 ±0.4	25.5 ±1.5	28.1 ±1.0	39.1 ±0.5	36.8 ±0.9	36.5 ±0.8	29.6 ±1.3	21.7 ±0.6	35.6 ±0.7	24.1 ±0.9	**40.4** ±0.5
	ARI	14.6 ±0.4	9.1 ±0.7	11.1 ±0.6	18.5 ±0.7	16.6 ±0.5	16.1 ±0.4	6.3 ±0.7	7.0 ±0.5	15.9 ±0.2	9.3 ±0.4	**19.7** ±0.5
CCV	ACC	22.0 ±1.4	30.0 ±0.7	32.6 ±1.0	34.5 ±0.4	25.9 ±1.0	22.6 ±0.2	22.7 ±1.3	25.9 ±1.0	18.1 ±0.6	24.6 ±1.1	**38.5** ±1.8
	NMI	22.1 ±1.0	27.4 ±0.7	29.0 ±0.3	28.8 ±0.7	24.7 ±0.6	21.7 ±0.2	24.3 ±1.3	24.7 ±0.6	14.7 ±0.8	25.4 ±0.6	**32.8** ±0.9
	ARI	11.7 ±1.5	13.2 ±0.5	14.7 ±0.4	16.2 ±0.4	12.3 ±0.6	9.0 ±0.1	7.7 ±1.0	12.3 ±0.6	5.4 ±0.3	11.1 ±0.5	**19.1** ±1.2

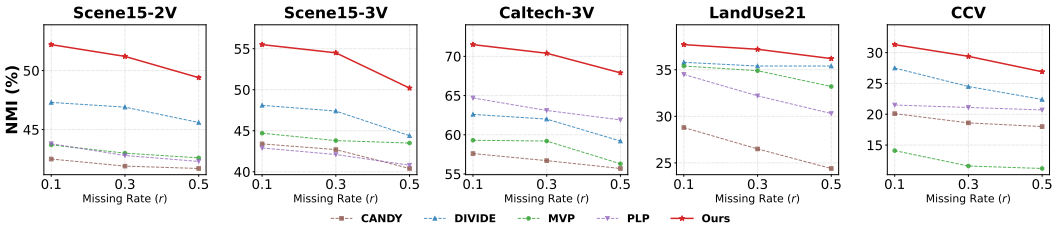


Figure 3: **Performance against view-missing scenarios.** We reported the clustering accuracy (NMI) on all five datasets (Scene15-2V, Scene15-3V, Caltech-3V, LandUse21, and CCV) under varying missing rates $r \in \{0.1, 0.3, 0.5\}$. EVE (red line) consistently outperforms state-of-the-art competitors, maintaining high stability even under severe missing rates.

performance against state-of-the-art competitors. Notably, compared to the runner-up, EVE boosts ACC, NMI, ARI by **7.2%, 7.6%, 7.4%** on Scene15-3V and **6.6%, 7.9%, 11.7%** on Caltech-3V. A deeper insight is obtained by comparing results on Scene15-2V and Scene15-3V. The latter introduces a third view that typically contains higher noise and lower discriminability. While most competitors achieve only marginal gains (e.g., DIVIDE improves by 3.3% in ACC) or struggle to utilize this additional information, our method achieves a substantial leap of **6.9%** (from 53.8% to 60.7%). This indicates that existing methods are sensitive to the quality of the added view, whereas our framework effectively "cleans" and enhances the weaker third view, enabling it to contribute complementary information rather than acting as noise. In addition, we conducted a statistical significance analysis in the Appendix B, Table 6, which further confirms that the performance gains achieved by our method are **statistically significant** on all evaluated datasets.

Robustness against incomplete multi-view scenarios. We conducted experiments with varying missing rates $r \in \{0.1, 0.3, 0.5\}$. As visualized in Fig. 3, EVE consistently outperforms state-of-the-art competitors in NMI across all five datasets. Detailed numerical comparisons (including ACC and ARI) are provided in Table 5 of the Appendix B. Notably, EVE exhibits exceptional robustness even under severe data incompleteness ($r = 0.5$). For instance, on Scene15-3V and Caltech-3V, it surpasses the runner-up by substantial margins of **5.8%** and **6.0%** in NMI, respectively. While performance naturally degrades as the missing rate increases, these significant gains suggest that the high-quality single-view representations cultivated by our framework effectively compensate for the absence of complementary information, maintaining structural integrity when views are missing.

4.3 Ablation Study

Each component significantly contributes to performance. We dissected the contribution of each component in Table 2. Using only the intra-view loss $\mathcal{L}_{\text{intra}}$ establishes a basic baseline. Introducing the fused-view supervision $\mathcal{L}_{\text{fuse}}$ yields considerable improvements (e.g., **+5.3%** ACC on Scene15-3V and **+12.1%** ACC on CCV), confirming the value of the fused global representation. On top of this, incorporating either the selective alignment ($\mathcal{L}_{\text{inter}}$) or neighborhood propagation (NP) brings further performance gains across datasets. Most importantly, integrating all modules yields the best overall

Table 2: Ablation study (%) of loss components and fused-guided neighborhood propagation (NP).

$\mathcal{L}_{\text{intra}}$	$\mathcal{L}_{\text{inter}}$	$\mathcal{L}_{\text{fuse}}$	NP	Scene15-2V			Scene15-3V			Caltech-3V			LandUse21			CCV			Avg.
				ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	
✓				46.6	45.7	30.0	50.4	46.1	31.8	74.1	62.9	58.4	32.3	38.4	17.9	25.1	19.8	9.0	39.2
✓		✓		49.1	45.8	31.7	55.7	51.4	38.7	77.4	66.2	61.4	**35.5**	40.3	19.6	37.2	31.3	17.9	43.9
✓		✓	✓	51.0	49.6	34.8	56.7	54.1	41.4	80.1	70.5	66.5	35.1	**41.4**	**20.5**	38.8	32.9	19.5	46.2
✓	✓	✓		51.0	49.7	34.5	57.8	54.1	41.0	80.0	71.3	67.3	34.0	38.4	18.6	37.3	32.3	18.1	45.7
✓	✓	✓	✓	**53.7**	**52.6**	**37.5**	**62.1**	**59.7**	**46.7**	**82.7**	**75.2**	**71.8**	35.1	40.4	20.2	**40.0**	**33.9**	**20.4**	**48.8**

performance on the majority of benchmarks. The full framework achieves **82.7%** ACC on Caltech-3V and **62.1%** on Scene15-3V, significantly surpassing the variants using any single auxiliary module. These consistent step-wise increments verify that feature-level alignment and structural refinement are complementary, and their combination is essential for achieving the optimal clustering results.

4.4 More Discussions

Enhancement of view-specific representations. To empirically verify that EVE improves individual view quality, we first visualized the NMI performance of view-specific representations in Fig. 4. As clearly illustrated by the radar charts, our method EVE (marked in **Red**) consistently outperforms all competitors across all available views, forming the largest coverage area on both Scene15-3V and Caltech-3V. To provide a more intuitive understanding of this improvement, we further visualized the feature distributions on the Scene15-3V dataset using t-SNE. Due to space limitations, we focus on View 3 (Fig. 5), which represents the most challenging view with high noise and ambiguity. As shown in Fig. 5(a), the independently trained single-view representation suffers from severe class overlap and blurred boundaries. While state-of-the-art methods (Fig. 5(b)-(c)) still struggle to disentangle the complex structure, our framework (Fig. 5(d)) achieves a remarkable transformation. The clusters become much more compact with distinct inter-class margins, visually confirming that our global guidance mechanism effectively purifies the noisy view. Full visualizations for all views are provided in the Appendix B, Table 10. This collective evidence strongly supports that by establishing a feedback loop with global guidance, our method does not merely fuse features but fundamentally rectifies the underlying manifold structure of individual views.

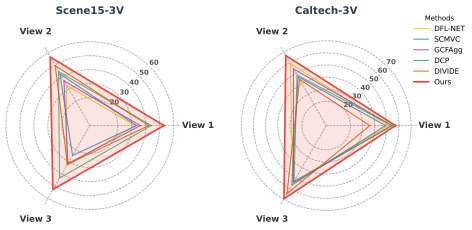


Figure 4: NMI (%) of individual views. EVE achieves the best performance.

Consistently better performance under systematic missing views and partial corruption. We conduct additional experiments on Caltech-3V under systematic missing-view and partial corruption settings. For systematic missing, 50% of the samples remain complete, while the remaining 50% consistently lack one specific view. For partial corruption, randomly selected views are degraded by Gaussian noise. As reported in Table 3, EVE consistently outperforms DIVIDE across all settings. In particular, EVE maintains stable performance across different missing-view patterns and achieves substantial improvements under partial corruption. These results demonstrate that EVE remains effective when views are systematically missing for a subset of samples or when available views are affected by noise.

Dominant views serve as reliable local anchors. A core premise of our selective alignment strategy is that peer-to-peer consistency should be guided by quality, rather than enforced blindly. Here, the fused representation acts as a global evaluator to identify the most confident "dominant view" (v^{dom})

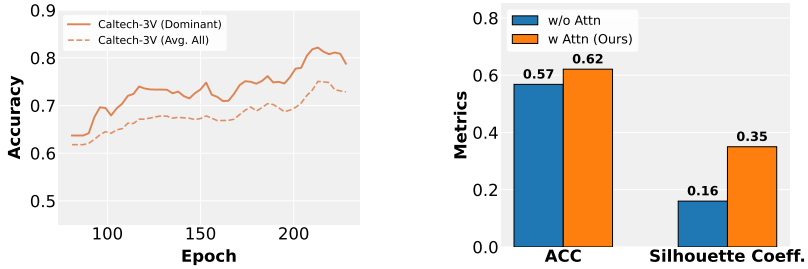
Table 3: Clustering performance (%) under systematic missing-view and partial corruption settings on Caltech-3V.

Setting	DIVIDE			EVE		
	ACC	NMI	ARI	ACC	NMI	ARI
Random Missing	73.0	58.8	54.7	**79.1**	**68.2**	**64.9**
View 1 Missing	71.1	57.9	51.9	**77.8**	**67.8**	**63.7**
View 2 Missing	71.6	59.1	52.9	**79.6**	**70.9**	**66.4**
View 3 Missing	71.1	58.2	52.1	**76.9**	**65.5**	**61.5**
Partial Corruption	69.4	56.8	51.3	**77.6**	**69.3**	**63.4**



(a) Single, NMI=19.0 (b) SCMVC, NMI=21.7 (c) DIVIDE, NMI=31.9 (d) Ours, NMI=52.8

Figure 5: T-SNE visualization(%) of the most challenging view (View 3) on the Scene15-3V.



(a) (b)

Figure 6: (a) Accuracy of the selected dominant views versus the average of all views. (b) ACC and Silhouette Coefficient (SC) comparisons between attention and concatenation fusion.

for each sample. To verify this, we monitored the quality of the identified dominant views during training. Fig. 6(a) illustrates the accuracy of the selected dominant views versus the average accuracy of all views on Caltech-3V. As observed, the "Dominant View" curve consistently lies above the "Average View" curve by a significant margin throughout the training process. This validates that our reliability estimation metric (Eq. 8), derived from the consistency between the global fusion and local predictions, effectively pinpoints the most discriminative view for each sample. By utilizing this dominant view as a high-quality **local anchor** for $\mathcal{L}_{\text{inter}}$, the model ensures that weaker views align with the most reliable peer rather than average or noisy features. This precise "strong-to-weak" knowledge transfer explains the substantial performance boost (e.g., **+2.6%** on Caltech-3V) observed when this module is activated.

Superiority of attention-based global consensus. To investigate the fusion strategy, we replaced the attention module with concatenation (denoted as "w/o Attn"). Results in Fig. 6(b) on Scene15-3V validate our design from **two perspectives**. First, regarding **framework validity**, even with simple concatenation, our method achieves **56.8%** ACC and **0.16** SC, surpassing most existing methods and confirming the effectiveness of our closed-loop framework. Second, regarding **fusion quality**, the attention mechanism acts as a quality filter, boosting ACC to **62.1%** and SC to **0.35** by constructing a cleaner teacher. **Moreover**, this mechanism enables native adaptation to varying view availability without architectural modification, ensuring intrinsic **compatibility with incomplete views**.

5 Conclusion

In this paper, we highlighted that existing deep multi-view clustering methods predominantly focus on optimizing the fused representation while often overlooking the enhancement of individual views. To address this limitation, we proposed Explicit single-View Enhancement (EVE) framework to explicitly rectify and optimize view-specific representations. By leveraging the reliable global consensus to refine individual views at both feature and structural levels, EVE significantly improves view-specific representations, which in turn boosts the global consensus, thereby establishing a mutually reinforcing feedback loop. Extensive empirical results validate that our approach sets a new state-of-the-art on standard benchmarks, achieving substantial average improvements of 5.9% in ARI, while remaining robust in incomplete multi-view clustering. Limitations include the current focus on random missing views, and extending to more extreme missing scenarios may further enhance the framework's generality. We hope this work inspires future research on the synergy between local view refinement and global fused representation learning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants U24A20322 and 62576094. This research work was also supported by the Big Data Computing Center of Southeast University.

References

[1] 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*, pages 16752–16761, 2023.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[3] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

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

[5] Jinrong Cui, Yuting Li, Han Huang, and Jie Wen. Dual contrast-driven deep multi-view clustering. *IEEE Transactions on Image Processing*, 33:4753–4764, 2024.

[6] Zhibin Dong, Siwei Wang, Jiaqi Jin, Xinwang Liu, and En Zhu. Cross-view topology based consistent and complementary information for deep multi-view clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19440–19451, 2023.

[7] Ran Eisenberg, Jonathan Svirsky, and Ofir Lindenbaum. COPER: Correlation-based permutations for multi-view clustering. In *The Thirteenth International Conference on Learning Representations*, 2025.

[8] Cong Feng, Ao Li, Haoyue Xu, Hailu Yang, and Xinwang Liu. Deep incomplete multiview clustering via local and global pseudo-label propagation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5):8684–8698, 2025.

[9] Xin Gao and Jian Pu. Deep incomplete multi-view learning via cyclic permutation of VAEs. In *The Thirteenth International Conference on Learning Representations*, 2025.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.

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

[12] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Ijcai*, pages 1753–1759, 2017.

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020.

[14] 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, pages 17305–17313, 2025.

[15] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

[16] Yuheng Jia, Hui Liu, Junhui Hou, Sam Kwong, and Qingfu Zhang. Multi-view spectral clustering tailored tensor low-rank representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 31:4784–4797, 2021.

[17] Yuheng Jia, Jianhong Cheng, Hui Liu, and Junhui Hou. Towards calibrated deep clustering network. In *The Thirteenth International Conference on Learning Representations*, 2025.

[18] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia*, page 29, 2011.

[19] Zhengxing Jiao, Yaxin Hou, Jun Ma, Yuhang Li, Ding Ding, Yuheng Jia, Hui Liu, and Junhui Hou. Samples are not equal: A sample selection approach for deep clustering. In *The Fourteenth International Conference on Learning Representations*, 2026.

[20] Feifei. Li and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 524–531, 2005.

[21] Feifei Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, page 178, 2004.

[22] Hanyang Li, Yuheng Jia, Hui Liu, and Junhui Hou. You can trust your clustering model: A parameter-free self-boosting plug-in for deep clustering. In *Advances in Neural Information Processing Systems*, volume 38, pages 14066–14091, 2025.

[23] Tao Li and Chris Ding. The relationships among various nonnegative matrix factorization methods for clustering. In *International Conference on Data Mining*, pages 362–371, 2006.

[24] Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, Zhaohua Yang, et al. Deep adversarial multi-view clustering network. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2952–2958, 2019.

[25] Zhixin Li, Yuheng Jia, Junhui Hou, et al. Learning from sample stability for deep clustering. In *Forty-second International Conference on Machine Learning*, 2025.

[26] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2023.

[27] Zhengzheng Lou, Ke Zhang, Yucong Wu, and Shizhe Hu. Super deep contrastive information bottleneck for multi-modal clustering. In *Forty-second International Conference on Machine Learning*, 2025.

[28] Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14193–14201, 2024.

[29] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

[30] Song Wu, Yan Zheng, Yazhou Ren, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, and Lifang He. Self-weighted contrastive fusion for deep multi-view clustering. *IEEE Transactions on Multimedia*, 26:9150–9162, 2024.

[31] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[32] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. GCFAgg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19863–19872, 2023.

[33] Xiaoqiang Yan, Zhixiang Jin, Fengshou Han, and Yangdong Ye. Differentiable information bottleneck for deterministic multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27425–27434, 2024.

[34] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1055–1069, 2023.

[35] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279, 2010.

[36] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[37] Changqing Zhang, Yeqing Liu, and Huazhu Fu. AE2-Nets: Autoencoder in autoencoder networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2577–2585, 2019.

Appendix Table of Contents

A. Experiment Settings 15

Dataset details and feature descriptions

Incomplete multi-view clustering settings

Implementation details

B. Additional Experiments 15

Detailed results on incomplete multi-view scenarios (Table 5)

Significance analysis (Table 6)

Computational efficiency analysis (Table 7)

More visualizations for each view (Figures 7, 8, 9)

Parameter sensitivity on propagation depth and self-loop weight (Table 8)

Ablation studies on alignment strategies and affinity matrix construction (Table 9)

Full t-SNE visualization results (Figure 10)

C. Verifying Single-View Enhancement: A Layer-wise Perspective 19

A Experiment Settings

Table 4: Statistics of the multi-view datasets.

Dataset	Views	Dimensions	#Samples	#Classes
Scene15-2V	2	20/59	4485	15
Scene15-3V	3	20/59/40	4485	15
Caltech-3V	3	40/254/928	1400	7
LandUse21	2	59/40	2100	21
CCV	3	4000/5000/5000	6773	20

Dataset details and feature descriptions. We provide a detailed description of the feature representations used for each multi-view dataset. The specific types of features corresponding to the dimensions reported in Table 4 are as follows: **Scene15-2V and Scene15-3V.** Following [34, 28, 8], View 1 (20-dim) corresponds to GIST features, and View 2 (59-dim) corresponds to LBP. For Scene15-3V, following [9, 7] the additional View 3 (40-dim) corresponds to Gabor features. **Caltech-3V.** Following [30, 14], View 1 (40-dim), View 2 (254-dim), and View 3 (928-dim) are extracted using Gabor filters, WM, and LBP, respectively. **LandUse21.** Following [26, 28, 8], View 1 (59-dim) and View 2 (40-dim) correspond to LBP and Gabor features, respectively. **CCV.** Following [30, 7], the three views are represented by hand-crafted features: View 1 (4,000-dim) uses MFCC for audio; View 2 (5,000-dim) uses SIFT for static visual content; and View 3 (5,000-dim) uses STIP for motion information.

Incomplete multi-view clustering settings. Given an incomplete multi-view dataset $\mathcal{X} = \{X^{(1)}, X^{(2)}, \ldots, X^{(V)}\}$, where each sample is described by V views, let $r \in [0, 1]$ denote the input missing rate, which determines the proportion of incomplete samples. Specifically, rN samples are randomly selected from the total N samples as incomplete instances. For each incomplete sample $x_i = \{x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(V)}\}$, every view may be missing independently. A binary indicator matrix $M \in \{0, 1\}^{N \times V}$ is introduced, where $M_{iv} = 1$ indicates that the v-th view of sample i is observed, and $M_{iv} = 0$ otherwise. For incomplete samples, available views are randomly dropped under the constraint $\sum_{v=1}^{V} M_{iv} \geq 1, \forall i$, ensuring that at least one view remains observable for each sample.

Implementation details. We implemented our framework on an NVIDIA V100 GPU using the PyTorch platform. For the network architecture, each view utilizes a 4-layer MLP encoder structured as Input–1024–1024–1024–128. Both the attention module and the projector shared an identical architecture, consisting of two fully connected layers with ReLU activation. Regarding optimization, we employed the Adam optimizer with a batch size of 1024 and a learning rate of 2×10^{-3}. The temperature parameter τ in contrastive losses was set to 0.5, and the RBF kernel bandwidth σ was fixed at 0.1. To ensure stable convergence, we adopted a progressive strategy over 200 epochs, initially optimizing $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{fuse}}$ before activating $\mathcal{L}_{\text{inter}}$ to avoid early noise. Subsequently, fused-guided neighborhood propagation was enabled for an additional 30 fine-tuning epochs to inject structural constraints. This curriculum ensures geometric rectification is applied only when the feature space is sufficiently discriminative. **Notably, all hyperparameters of our method are fixed across all datasets**, and sensitivity analyses are provided in the Appendix B, Table 8.

B Additional Experiments

Detailed results on incomplete multi-view scenarios. Table 5 provides the comprehensive numerical results (ACC, NMI, and ARI) across five datasets under varying missing rates ($r \in \{0.1, 0.3, 0.5\}$). Complementing the visual trends discussed in the main text, these detailed metrics confirm that our method achieves consistent superiority across all evaluation criteria. Notably, our framework exhibits exceptional robustness even under severe data incompleteness ($r = 0.5$), maintaining a significant performance margin over the runner-up (e.g., a lead of approximately 6% in ACC on Scene15-3V). This further substantiates that the enhanced discriminability of individual views effectively mitigates the impact of missing cross-view information.

Significance analysis. We conducted a comprehensive statistical significance test against DIVIDE [28] and performed **Paired Sample t-tests** on ACC, NMI, and ARI. All comparisons are based

Table 5: **Clustering performance (%) on incomplete multi-view scenarios under different missing rates.** All results are reported over 5 runs with different random seeds. The best results are highlighted in **bold**, and the second-best results are in underlined.

Dataset	Method	r = 0.1			r = 0.3			r = 0.5		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Scene15-2V	CANDY	40.8 $\pm_{0.6}$	42.5 $\pm_{0.5}$	25.7 $\pm_{0.5}$	40.7 $\pm_{1.3}$	41.9 $\pm_{0.9}$	25.4 $\pm_{1.2}$	39.4 $\pm_{0.9}$	41.7 $\pm_{0.5}$	25.2 $\pm_{0.3}$
	DIVIDE	48.6 $\pm_{0.4}$	47.3 $\pm_{0.4}$	31.0 $\pm_{0.4}$	48.3 $\pm_{0.4}$	46.9 $\pm_{0.5}$	30.8 $\pm_{0.3}$	46.7 $\pm_{0.6}$	45.6 $\pm_{0.4}$	29.4 $\pm_{0.5}$
	MVP	44.9 $\pm_{0.4}$	43.7 $\pm_{0.2}$	26.3 $\pm_{0.1}$	45.0 $\pm_{0.5}$	43.0 $\pm_{1.0}$	26.1 $\pm_{0.9}$	44.6 $\pm_{0.4}$	42.6 $\pm_{0.3}$	25.7 $\pm_{0.3}$
	PLP	39.9 $\pm_{1.0}$	43.8 $\pm_{0.2}$	26.3 $\pm_{0.2}$	40.9 $\pm_{0.9}$	42.8 $\pm_{0.4}$	25.7 $\pm_{0.7}$	40.7 $\pm_{1.0}$	42.3 $\pm_{0.7}$	25.5 $\pm_{1.0}$
	EVE (Ours)	**53.0** $\pm_{1.2}$	**52.2** $\pm_{0.9}$	**37.3** $\pm_{1.3}$	**52.9** $\pm_{0.9}$	**51.2** $\pm_{0.5}$	**36.6** $\pm_{1.0}$	**50.2** $\pm_{1.7}$	**49.4** $\pm_{0.9}$	**34.0** $\pm_{1.2}$
Scene15-3V	CANDY	41.0 $\pm_{0.9}$	43.4 $\pm_{0.4}$	26.9 $\pm_{0.4}$	41.9 $\pm_{2.4}$	42.7 $\pm_{1.0}$	27.0 $\pm_{1.4}$	39.0 $\pm_{1.1}$	40.4 $\pm_{0.6}$	25.2 $\pm_{0.6}$
	DIVIDE	51.3 $\pm_{1.2}$	48.1 $\pm_{0.9}$	34.2 $\pm_{1.4}$	51.5 $\pm_{1.1}$	47.4 $\pm_{0.4}$	34.0 $\pm_{1.0}$	48.7 $\pm_{0.4}$	44.4 $\pm_{0.4}$	30.7 $\pm_{0.6}$
	MVP	48.6 $\pm_{0.5}$	44.7 $\pm_{0.6}$	30.0 $\pm_{0.6}$	48.3 $\pm_{0.8}$	43.8 $\pm_{0.7}$	29.2 $\pm_{0.9}$	47.6 $\pm_{0.6}$	43.5 $\pm_{0.6}$	28.9 $\pm_{0.6}$
	PLP	39.0 $\pm_{0.2}$	42.9 $\pm_{0.3}$	26.5 $\pm_{0.3}$	38.8 $\pm_{0.4}$	42.1 $\pm_{0.3}$	25.9 $\pm_{0.3}$	38.4 $\pm_{0.3}$	40.8 $\pm_{0.4}$	25.1 $\pm_{0.5}$
	EVE (Ours)	**58.3** $\pm_{1.7}$	**55.5** $\pm_{1.4}$	**42.2** $\pm_{1.4}$	**58.4** $\pm_{1.0}$	**54.5** $\pm_{0.8}$	**42.3** $\pm_{0.9}$	**54.6** $\pm_{1.5}$	**50.2** $\pm_{0.5}$	**37.6** $\pm_{1.2}$
Caltech-3V	CANDY	70.7 $\pm_{3.0}$	57.6 $\pm_{3.0}$	51.4 $\pm_{4.0}$	68.9 $\pm_{2.7}$	56.7 $\pm_{3.3}$	50.8 $\pm_{3.8}$	71.3 $\pm_{2.9}$	55.7 $\pm_{2.4}$	51.1 $\pm_{4.0}$
	DIVIDE	74.5 $\pm_{0.7}$	62.6 $\pm_{1.2}$	57.9 $\pm_{1.6}$	74.1 $\pm_{2.5}$	62.0 $\pm_{1.8}$	57.5 $\pm_{2.1}$	73.5 $\pm_{2.5}$	59.2 $\pm_{2.1}$	55.1 $\pm_{3.1}$
	MVP	75.8 $\pm_{1.1}$	59.3 $\pm_{1.1}$	55.4 $\pm_{1.4}$	75.3 $\pm_{0.7}$	59.2 $\pm_{0.3}$	54.9 $\pm_{0.7}$	73.8 $\pm_{0.7}$	56.3 $\pm_{1.0}$	52.0 $\pm_{0.9}$
	PLP	72.4 $\pm_{0.8}$	64.7 $\pm_{0.8}$	56.4 $\pm_{0.7}$	72.1 $\pm_{1.4}$	63.1 $\pm_{1.8}$	55.4 $\pm_{1.6}$	70.4 $\pm_{0.5}$	61.9 $\pm_{0.8}$	53.7 $\pm_{0.6}$
	EVE (Ours)	**80.8** $\pm_{2.3}$	**71.5** $\pm_{1.3}$	**67.0** $\pm_{2.1}$	**79.5** $\pm_{1.8}$	**70.4** $\pm_{0.9}$	**65.7** $\pm_{1.2}$	**79.3** $\pm_{2.7}$	**67.9** $\pm_{1.9}$	**64.9** $\pm_{2.5}$
LandUse21	CANDY	27.4 $\pm_{1.0}$	28.8 $\pm_{0.7}$	12.4 $\pm_{1.0}$	25.6 $\pm_{1.0}$	26.5 $\pm_{1.4}$	11.3 $\pm_{1.0}$	24.4 $\pm_{0.6}$	24.4 $\pm_{1.0}$	10.2 $\pm_{1.0}$
	DIVIDE	31.4 $\pm_{0.8}$	35.8 $\pm_{0.7}$	16.5 $\pm_{0.4}$	31.8 $\pm_{0.6}$	35.4 $\pm_{0.9}$	16.6 $\pm_{0.5}$	30.7 $\pm_{0.5}$	35.4 $\pm_{0.6}$	16.3 $\pm_{0.4}$
	MVP	30.9 $\pm_{0.6}$	35.4 $\pm_{0.5}$	15.9 $\pm_{0.5}$	31.0 $\pm_{0.2}$	34.9 $\pm_{0.7}$	15.7 $\pm_{0.2}$	30.5 $\pm_{0.4}$	33.2 $\pm_{0.2}$	15.0 $\pm_{0.3}$
	PLP	29.2 $\pm_{0.4}$	34.5 $\pm_{1.0}$	14.8 $\pm_{0.5}$	28.1 $\pm_{0.9}$	32.2 $\pm_{0.9}$	13.8 $\pm_{1.0}$	27.8 $\pm_{0.2}$	30.3 $\pm_{0.3}$	13.0 $\pm_{0.2}$
	EVE (Ours)	**32.4** $\pm_{0.5}$	**37.7** $\pm_{0.7}$	**17.4** $\pm_{0.5}$	**31.9** $\pm_{0.3}$	**37.2** $\pm_{0.7}$	**17.4** $\pm_{0.6}$	**30.9** $\pm_{0.4}$	**36.2** $\pm_{0.6}$	**16.9** $\pm_{0.3}$
CCV	CANDY	23.9 $\pm_{3.8}$	20.1 $\pm_{3.6}$	8.0 $\pm_{3.1}$	23.1 $\pm_{4.7}$	18.6 $\pm_{4.1}$	7.8 $\pm_{2.6}$	22.4 $\pm_{3.8}$	18.0 $\pm_{3.0}$	7.3 $\pm_{2.4}$
	DIVIDE	33.0 $\pm_{0.4}$	27.5 $\pm_{0.3}$	14.7 $\pm_{0.3}$	30.7 $\pm_{0.7}$	24.5 $\pm_{0.3}$	12.5 $\pm_{0.5}$	27.7 $\pm_{1.4}$	22.4 $\pm_{0.6}$	10.9 $\pm_{0.7}$
	MVP	17.8 $\pm_{0.7}$	14.1 $\pm_{1.3}$	5.3 $\pm_{0.8}$	16.3 $\pm_{0.6}$	11.6 $\pm_{1.0}$	4.3 $\pm_{0.1}$	15.6 $\pm_{0.7}$	11.2 $\pm_{0.9}$	3.4 $\pm_{0.4}$
	PLP	23.0 $\pm_{0.4}$	21.5 $\pm_{0.2}$	9.3 $\pm_{0.2}$	23.4 $\pm_{0.5}$	21.1 $\pm_{0.2}$	9.2 $\pm_{0.2}$	22.6 $\pm_{0.2}$	20.7 $\pm_{0.4}$	8.9 $\pm_{0.2}$
	EVE (Ours)	**36.6** $\pm_{1.7}$	**31.3** $\pm_{1.5}$	**18.0** $\pm_{1.2}$	**35.5** $\pm_{0.8}$	**29.4** $\pm_{0.4}$	**16.5** $\pm_{0.5}$	**32.6** $\pm_{0.6}$	**26.9** $\pm_{0.5}$	**14.3** $\pm_{0.3}$

on 5 independent runs using identical random seeds. For other baselines, formal significance testing was deemed unnecessary due to the substantial performance margins, where the gaps are sufficiently large to guarantee statistical significance. The results are summarized in Table 6.

Table 6: **Paired t-test results (p-values) between our method EVE and DIVIDE.** Significance levels: * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).

Dataset	p-value (Significance)		
	ACC	NMI	ARI
Scene15-2V	2.0×10^{-4} (***)	7.3×10^{-4} (***)	9.0×10^{-5} (***)
Scene15-3V	1.7×10^{-4} (***)	5.2×10^{-4} (***)	2.5×10^{-4} (***)
Caltech-3V	1.9×10^{-5} (***)	1.7×10^{-4} (***)	1.0×10^{-4} (***)
LandUse21	0.015 (*)	0.013 (*)	6.3×10^{-4} (***)
CCV	0.004 (**)	1.5×10^{-4} (***)	0.001 (**)

Computational efficiency analysis. Table 7 compares the memory usage and training time across five datasets. Our method maintains a competitive efficiency profile, consuming comparable resources to DIVIDE while being significantly faster than CANDY. Considering the substantial performance improvements achieved (e.g., an average 5.9% boost in ARI), our method offers a highly favorable trade-off between effectiveness and efficiency.

Table 7: Memory consumption (MB) and running time (seconds) comparison.

Dataset	Metric	CANDY	DIVIDE	OURS
Scene15-2V	Memory (MB)	734	660	694
	Time (s)	388	159	192
Scene15-3V	Memory (MB)	906	774	822
	Time (s)	647	176	225
Caltech-3V	Memory (MB)	936	802	832
	Time (s)	238	132	137
LandUse21	Memory (MB)	736	664	696
	Time (s)	232	141	158
CCV	Memory (MB)	1336	1096	1148
	Time (s)	1252	346	465

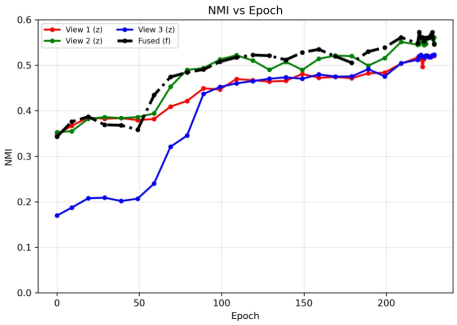


Figure 7: Dynamic co-evolution of each view and the fused performance over training epochs on Scene15-3V.

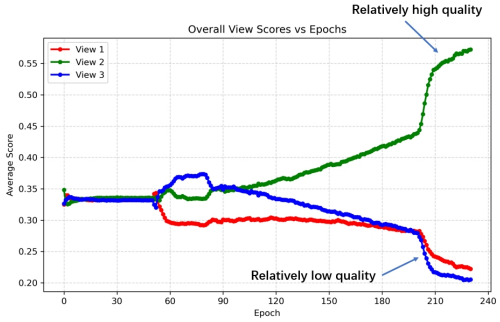


Figure 8: Attention scores of different views over training epochs on Scene15-3V.

Our framework consistently improves individual view quality while adaptively weighting views during fusion. As shown in Fig. 7, all views achieve noticeable performance gains throughout training, demonstrating that our method enhances each view-specific representation rather than only optimizing the fused result. This effect is particularly significant for the initially weak View 3, whose performance improves substantially and approaches that of stronger views, confirming the effectiveness of our enhancement mechanism. And Fig. 8 further shows that the fusion process is quality-aware. High-quality views receive progressively larger attention weights and contribute more to the fused representation, while lower-quality views are assigned smaller weights over time.

Sample-level attention further reveals the adaptive reliability modeling of our framework. As shown in Fig. 9, the attention distribution evolves from relatively uniform scores in early epochs to a highly differentiated pattern in later stages. By the final epoch, View 2 consistently receives the highest scores across most samples, while Views 1 and 3 are assigned substantially lower weights. This indicates that the model progressively identifies the most reliable view at the sample level and allocates contribution accordingly. Such behavior demonstrates that our framework is not limited to global view weighting, but also performs fine-grained, sample-specific adaptation. Reliable views dominate when their information is beneficial, whereas weaker views are automatically down-weighted to avoid degrading the fused representation.

Parameter sensitivity on propagation depth and self-loop weight. Table 8 reports the sensitivity of our method to the propagation depth $step$ and the self-loop coefficient α. For different parameter settings, the clustering performance remains relatively stable on both Scene15-3V and Caltech-3V, with no abrupt performance degradation. Although the optimal values differ slightly across datasets, the performance variations are small, indicating that the proposed method does not rely on carefully tuned hyperparameters. Overall, these results suggest that our framework is robust to reasonable choices of propagation depth and self-loop weight.

Table 8: Parameter analysis on Scene15-3V and Caltech-3V datasets w.r.t. (a) propagation depth $step$ and (b) self-loop coefficient α.

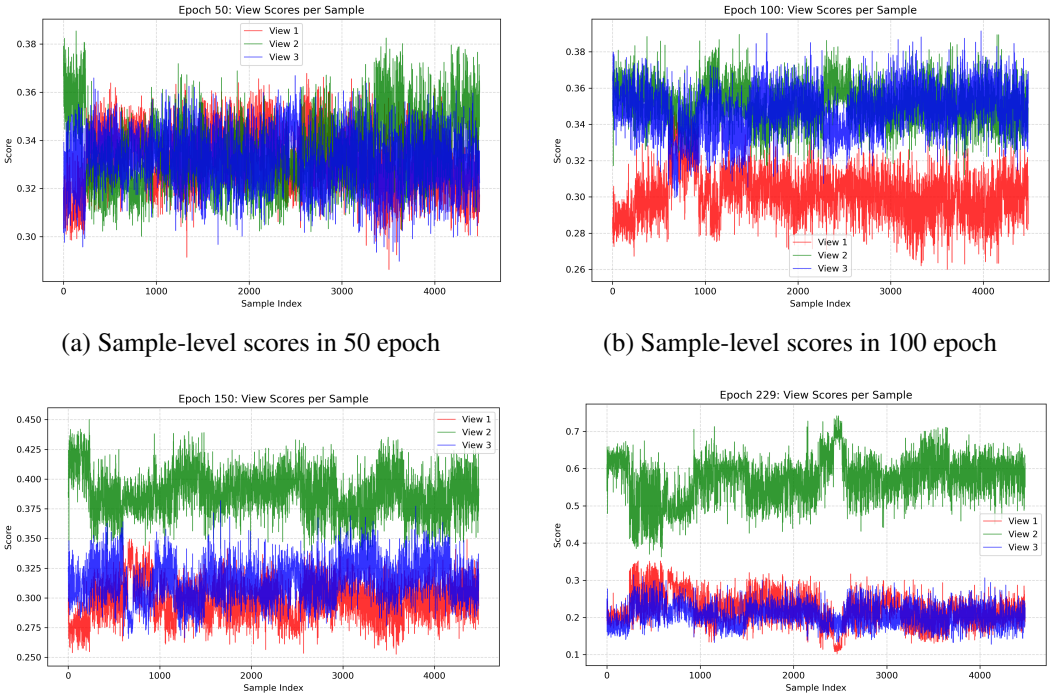
(a) Propagation depth $step$.

Dataset	$step$	1	3	5	7	9
Scene15-3V	ACC	57.0	59.1	**62.1**	59.0	58.0
	NMI	55.5	57.3	**59.7**	58.4	58.1
Caltech-3V	ACC	81.2	81.4	82.7	82.8	**83.0**
	NMI	73.7	71.5	75.2	74.8	**75.4**

(b) Self-loop coefficient α.

Dataset	α	0	0.25	0.5	0.75	1
Scene15-3V	ACC	37.4	58.5	**62.1**	57.2	57.7
	NMI	51.1	56.8	**59.7**	56.0	54.4
Caltech-3V	ACC	80.9	82.6	**82.7**	81.0	80.0
	NMI	75.8	74.8	**75.2**	71.4	71.3

Ablation studies on alignment strategies and affinity matrix construction. We conducted a comprehensive ablation study to validate the effectiveness of our proposed alignment and structural guidance mechanisms, as summarized in Table 9. Regarding alignment strategies (Part a), the standard Full Alignment" causes notable performance drops (e.g., 14.4% on Caltech-3V) due to noise propagation from unreliable views. In contrast, our **Selective Alignment** prevents negative transfer by dynamically distilling knowledge only from the most confident dominant view." Regarding



(a) Sample-level scores in 50 epoch

(b) Sample-level scores in 100 epoch

(c) Sample-level scores in 150 epoch

(d) Sample-level scores in the last epoch

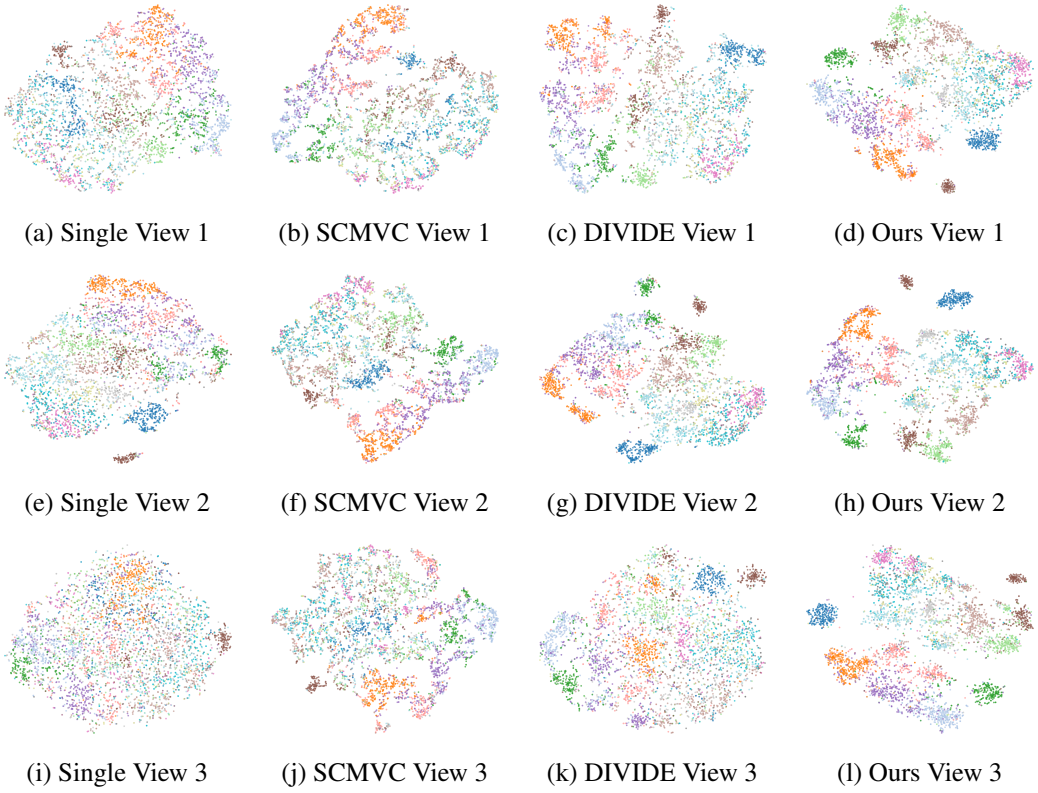
Figure 9: Sample-level attention scores over training epochs on Scene15-3V. Higher attention scores indicate that the corresponding view contributes more to the sample.

Table 9: **Comprehensive ablation study on key components.** (a) **Cross-view alignment strategies**, comparing different alignment mechanisms; (b) **Affinity matrix W construction strategies**, investigating different ways to define structural guidance. All results are reported in percentage (%).

Method Variant	**Scene15-2V**			**Scene15-3V**			**Caltech-3V**		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
(a) Comparison of Cross-view Alignment Strategies									
No Alignment (w/o $\mathcal{L}_{\text{inter}}$)	51.0	49.6	34.8	56.7	54.1	41.4	80.1	70.5	66.5
Full Alignment	47.7	48.6	34.0	58.0	54.7	42.9	68.3	60.2	50.9
Selective Alignment (Ours)	**53.7**	**52.6**	**37.5**	**62.1**	**59.7**	**46.7**	**82.7**	**75.2**	**71.8**
(b) Comparison of Affinity Matrix Construction Strategies									
Self-Affinity ($\mathbf{Z}^{(v)} \cdot \mathbf{Z}^{(v)\top}$)	52.7	50.4	36.5	57.6	55.4	42.5	82.4	74.8	71.4
Fused-Affinity ($\mathbf{Z}^{(\text{fuse})} \cdot \mathbf{Z}^{(\text{fuse})\top}$)	53.1	48.9	35.1	60.0	56.9	44.7	81.9	72.0	68.9
Cross-space Affinity (Ours)	**53.7**	**52.6**	**37.5**	**62.1**	**59.7**	**46.7**	**82.7**	**75.2**	**71.8**

structural guidance (Part b), our **Cross-space Affinity** outperforms baselines relying solely on local (Self-Affinity) or global (Fused-Affinity) graphs. By bridging view-specific features with the global consensus, it effectively rectifies local structural defects for more accurate neighborhood propagation.

Full visualization results. Fig. 10 presents the complete t-SNE visualizations for all three views on the Scene15-3V dataset, serving as a comprehensive supplement to the analysis in the main text. Consistent with the observations on View 3, our method (rightmost column) achieves the clearest cluster separation and most compact structures across the other views (View 1 and View 2) as well. In contrast, competitive baselines like SCMVC and DIVIDE still exhibit varying degrees of class overlap. This confirms that our explicit single-view enhancement strategy is universally effective across different modalities, ensuring that every view benefits from the global guidance.



(a) Single View 1 (b) SCMVC View 1 (c) DIVIDE View 1 (d) Ours View 1

(e) Single View 2 (f) SCMVC View 2 (g) DIVIDE View 2 (h) Ours View 2

(i) Single View 3 (j) SCMVC View 3 (k) DIVIDE View 3 (l) Ours View 3

Figure 10: **T-SNE visualization of feature distributions on the *Scene15-3V* dataset.** Each row corresponds to a specific view (View 1, View 2, and View 3 from top to bottom). The columns represent the feature distributions learned by **Single View**, **SCMVC**, **DIVIDE**, and **Ours**, respectively. It can be observed that our method (rightmost column) achieves clearer separation and more distinct cluster structures compared to other baselines across all views.

C Verifying Single-View Enhancement: A Layer-wise Perspective

A key objective of our framework is to explicitly enhance the representation quality of individual views at the encoder level. To verify this, we investigated whether the performance gains in multi-view clustering arise from genuinely improved encoders or predominantly from the alignment in the projection space. We conducted a layer-wise comparison on the Caltech-3V and Scene15-3V datasets, analyzing the NMI scores of **Shallow Features** (Encoder outputs $\mathbf{Z}$, marked as circles) versus **Deep Features** (Projector outputs $\mathbf{P}$, marked as squares).

Limited enhancement in existing methods. As illustrated in Fig 11, while state-of-the-art competitors (e.g., DIVIDE, SCMVC) generally outperform the **Independent Baseline** (dashed line), their improvements at the shallow encoder level are often marginal or inconsistent. For instance, on Scene15-3V, although some methods show gains over the baseline, their shallow representations still struggle to capture clear cluster structures (e.g., hovering below 45% NMI). This suggests that while existing fusion mechanisms provide some supervision to the encoders, they do not fully exploit the potential of single-view representation learning, relying heavily on the projection head ($\mathbf{P}$) to achieve the final clustering performance.

Superior enhancement by our framework. In contrast, our method (marked as stars) achieves substantial and consistent improvements across all views. Most notably, our **Shallow Features ($\mathbf{Z}$)** exhibit a significant performance leap, consistently surpassing both the baseline and the shallow features of all competitors by a large margin. On the challenging View 3 of Scene15-3V, our shallow feature performance ($\approx 53\%$) is strikingly higher than the best competitor's shallow result. This empirical evidence confirms that our proposed explicit enhancement strategy effectively revitalizes the backbone encoders, creating a significantly stronger foundation for multi-view fusion.

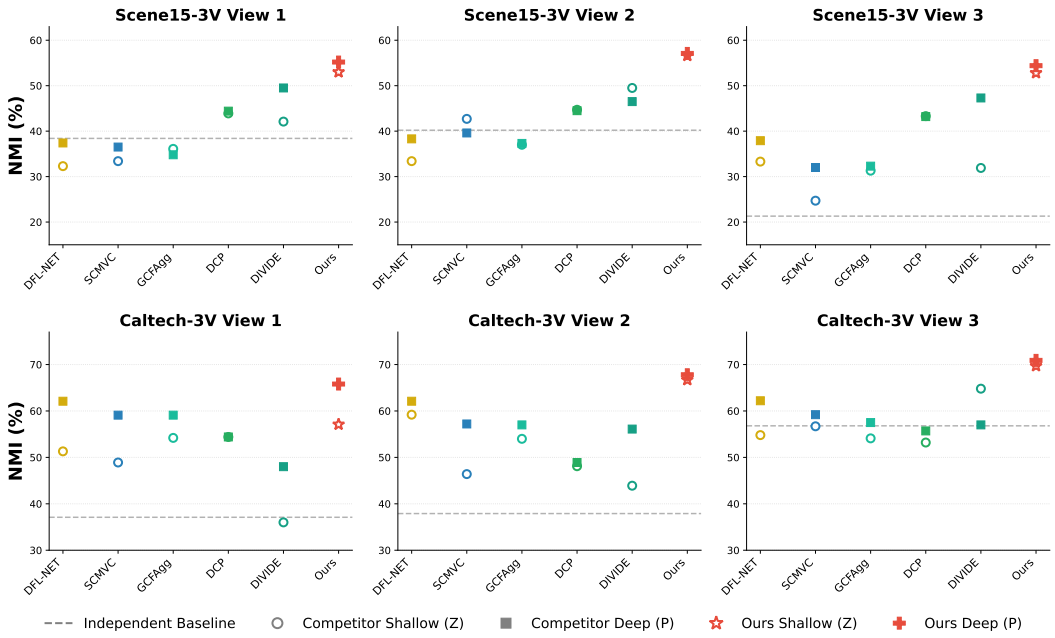


Figure 11: Layer-wise performance of individual view representations.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the motivation, proposed method, and main findings. They accurately reflect the paper's contributions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses limitations through experiments and conclusions.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: This paper is experimental and does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: All implementation details, training settings, and evaluation protocols are provided in the main paper and appendix.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: Experiments use public datasets, and the link to code is provided in the supplemental material.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of experimental settings in both the main text and supplemental material.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The paper includes multiple experiments with mean, standard, and statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper provides details on the compute resources used for the experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The work does not involve data or models with a high risk for misuse, so no specific safeguards are necessary.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All external assets (e.g., code, data, models) used in the paper are properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The source code is included in the Supplementary Material.

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The paper uses LLMs only for writing, editing, and formatting purposes and does not affect the core methodology or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.