
Diversity-oriented Deep Multi-modal Clustering

Yanzheng Wang[#], Xin Yang[#], Yujun Wang, Shizhe Hu^{*}, Mingliang Xu^{*}

School of Computer and Artificial Intelligence, Zhengzhou University, China

Abstract

Deep multi-modal clustering (DMC) aims to explore the correlated information from different modalities to improve the clustering performance. Most existing DMCs attempt to investigate the consistency or/and complementarity information by fusing all modalities, but this will lead to the following challenges: 1) Information conflicts between modalities emerge. 2) Information-rich modalities may be weakened. To address the above challenges, we propose a diversity-oriented deep multi-modal clustering (DDMC) method, where the core is dominant modality enhancement instead of multi-modal fusion. Specifically, we select the modality with the highest average silhouette coefficient as the dominant modality, then learn the diversity information between the dominant modality and the remaining ones with diversity learning, and finally enhance the dominant modality for clustering. Extensive experiments show the superiority of the proposed method over several compared DMC methods. To our knowledge, this is the first work to perform multi-modal clustering by enhancing the dominant modality instead of fusion.

1 Introduction

Deep multi-modal clustering (DMC) aims to integrate data from multiple modalities (e.g., image, text, audio, etc.) to classify the data through unsupervised learning. DMC combines the feature extraction ability of deep learning and the clustering idea of unsupervised learning, which is an important direction of multi-modal data analysis at present, and has achieved excellent performance in many fields, such as the medical field [1–3], autonomous driving and intelligent transportation [4, 5], and recommendation systems [6–8].

Related Works. At present, almost all DMCs cluster by fusion method. Based on different fusion stages, DMCs can be roughly divided into the following three categories: 1) Feature-level fusion: these methods [9–13] extract the features of different modalities and connect them into a single high-dimensional feature vector, which is used as a single input for clustering or feature learning. For example: Zhou and Shen [10] propose a method that first use an adversarial regularizer to align modalities, and then perform an attention fusion on all modalities, so as to quantify the importance of different modalities. 2) Decision-level fusion: these methods [14–16] first obtain independent modeling of each modality, obtain their own output or clustering results, and then fuse these results in the final stage. For example, Meng et al.[14] conduct the features of different modalities in their respective clustering layers for depression regression training, and linearly weighted sums the prediction results of each modality to obtain the fusion result as the final output result. 3) Mixed-level fusion: these methods [17–19] combine the output of a single mode prediction through feature-level fusion. For example, Morales et al.[17] train a separate model for each modality, then get the predictions for each modality, and finally train a new model on these new vectors to output the final prediction.

[#]Equal contribution

^{*}Corresponding author (ieshezhehu@gmail.com)

Motivations. Although the above methods have achieved excellent performance, due to the use of different levels of fusion, the following challenges arise: 1): Inconsistency of multi-modal data: different modalities may provide conflicting or inconsistent information about the same thing. 2): Weight assignment and modalities importance: modalities fusion can cause some informative high-quality modalities to be forced to align with low-quality high-noise modalities [20]. To address the above challenges, we drew inspiration from [21]. As shown in Figure 1(a), people perceive the world in the process of the world about 83% comes from the vision, and the remaining ways (such as smell, touch, taste and hearing) add up to only 17%. However, if the information obtained by the other ways is supplemented with the information obtained by the visual way, a more complete understanding will be obtained. Taking driving as an example, most of the information on the road condition comes from the vision, if at the same time through the radio, speakers and other information, we will have a more correct judgment on the road condition. Inspired by this, we propose a new multi-modal clustering method in Figure 1(b), in which modality 3 is the dominant modality. The diversity learning approach is employed to extract the diversity information between the dominant modality and the remaining modalities (modality 1, modality 2) and then the diversity information is spliced to the dominant modality for enhancement. Finally, the enhanced dominant modality is clustered.

Contributions. In this paper, we propose a diversity-oriented deep multi-modal clustering (DDMC) method. Specifically, our innovation lies in proposing a dominant modality enhancement strategy that ‘enhancement instead of fusion’. In this method, we first select an informative modality as the dominant modality, and subsequently employ diversity learning to extract the diversity information from other modalities from the dominant modality. Finally, we concatenate these dissimilarity information to the dominant modality and perform the final clustering output through the clustering module. Compared with the latest DMCs method, our method achieves significant performance improvement on benchmark datasets, which verifies its effectiveness and advantages. The main contributions of our work can be summarised as follows:

- Ours is the first work to investigate multi-modal clustering by enhancing dominant modalities rather than fusing modalities.
- We propose a diversity-oriented deep multi-modal clustering method by dominant modality enhancement rather than modality fusion, which can maximally retain important information in the raw modality and has the advantages of both single-modal and multi-modal clustering.
- We can effectively enhance the dominant modality by simultaneously mining the diversity information of the remaining modality relative to the dominant modality at both the feature-level and the cluster-level. The results on multiple multi-modal datasets can validate the effectiveness of DDMC.

2 Methodology

A multi-modal dataset $X = \{X^1, \dots, X^m, \dots, X^M\}$ contains N samples of M modalities, where $X^m \in \mathbb{R}^{N \times d_m}$ denotes the samples of dimension d_m from the m -th modality. Our purpose is to correctly divide N samples into K clusters by learning information between multiple modalities. To provide a more intuitive overview of the proposed method, we present it in Figure 2. Taking three modalities as an example, the dominant modality (modality 2 is shown in the figure) is first determined through the selection of dominant modality module. Subsequently, each modality is

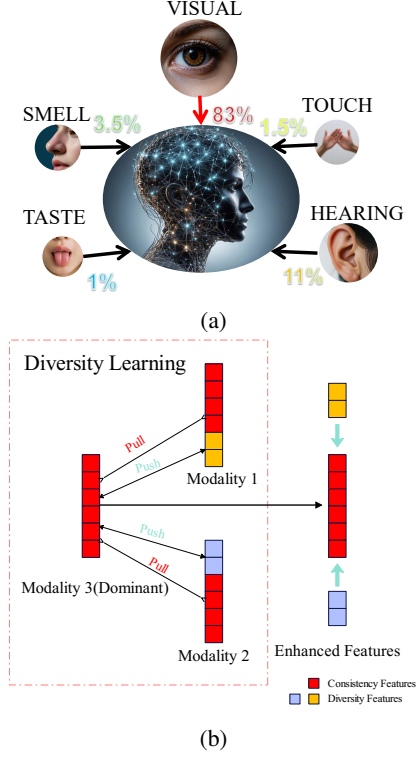


Figure 1: (a) The way humans perceive the world and its proportion. (b) The general idea of our method. Inspired by multi-modal assisted vision in human perception of the world.

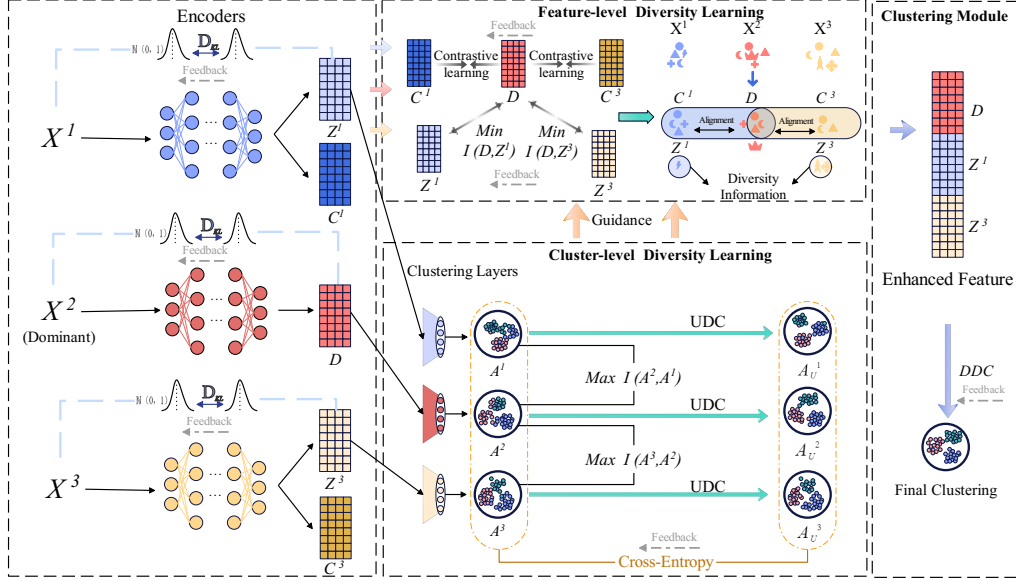


Figure 2: Details of the proposed framework. Firstly, the dominant modality feature D^{dm} , consistency features $\{C^m\}_{m \neq dm}^M$ and diversity features $\{Z^m\}_{m \neq dm}^M$ are obtained through modality specific encoders. Then, through feature-level diversity learning and cluster-level diversity learning, the diversity information of the remaining modality is learned. Finally, the diversity information is enhanced for the dominant modality, and the final clustering is completed through the clustering module. Among them, UDC is Uniform Distribution Constraint, which forces the edge distribution of each cluster to be as close to uniform as possible. DDC is Deep Divergence-based Clustering.

modeled with its original features through independent variational encoders. Next, diversity learning is performed at both the feature-level and the cluster-level, and the diversity information obtained from the non-dominant modalities is used to enhance the representation of the dominant modality. Finally, the final clustering is performed based on the enhanced dominant modality features to obtain the clustering results.

2.1 Motivation: Modality Fusion Leads to Lower Clustering Efficiency

In order to learn the information between modality, researchers capture the connections of heterogeneous modality by fusing each modality. However, this also brings corresponding problems: unfair weights in the fusion can lead to the degradation of the results. Specifically, we illustrate it on the popular multi-modal dataset Flickr [22], as shown in Figure 3. We conducted 20 iterations of K -means [23] clustering for each modality and the fusion modality, representing the quality of individual modality and the fusion modality through its unsupervised accuracy. As shown in the figure, it can be clearly seen that modality 3 belongs to the high-quality modality, while modality 1 and 2 belong to the low-quality modality. Due to the unequal relationship between the weights allocated to each modality and the quality in the fusion, the clustering effect of the fused modality is worse. However, these low-quality modality are not completely useless, we hope to select a high-quality modality as the dominant modality, and then learn useful information from the remaining modality to enhance the dominant modality. Therefore, unfair alignment of high-quality modality to low-quality modality can be avoided.

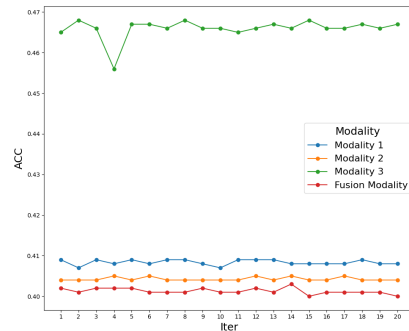


Figure 3: The accuracy rates of each modality and the fused modality after K -means clustering.

2.2 Selection of Dominant Modality

We decide that the dominant modality method is prior knowledge or the Silhouette coefficient (SI) fraction. Prior knowledge is that the modality weights in the dataset have been measured in previous work, or the importance of certain modality in some datasets is obvious. For these important modalities, the dominant modality is selected. For example, the importance weights of each modality are clearly given in [10]. In the absence of reliable prior information, the SI fraction is adopted to determine the dominant modality. SI is a metric used to evaluate the quality of clustering results. It combines the intra-cluster closeness and inter-cluster separation, and calculates a SI for each data point to measure its similarity to its own cluster and its nearest neighbor cluster. In the raw features, the clustering results are calculated by the K -means method, and then the SI score of each modality can be derived. For each sample x_i^m , the average SI of each modality is calculated as follows:

$$SI(X^m) = \frac{1}{N} \sum_{x_i^m \in X^m} SI(x_i^m), \quad (1)$$

where:

$$SI(x_i^m) = \frac{b(x_i^m) - a(x_i^m)}{\max\{b(x_i^m), a(x_i^m)\}}, \quad (2)$$

where $a(x_i^m) = \frac{1}{|C_k^m|} \sum_{x_j^m \in C_k^m, x_i^m \neq x_j^m} d(x_i^m, x_j^m)$ is to calculate the average distance between this sample point and all other points in the same cluster, and $b(x_i^m) = \min_{l \in [1, K], l \neq k} \frac{1}{|C_l^m|} \sum_{x_j^m \in C_l^m} d(x_i^m, x_j^m)$ is to calculate the average distance between this sample point and all points in the nearest neighboring cluster. $|C_l^m|$ is the number of all sample points in the cluster where x is located, and $d(\cdot)$ is the distance between the two sample points.

The modality with the highest SI is the dominant modality X^{dm} is given by:

$$X^{dm} = \max \{SI(X^1), SI(X^2), \dots, SI(X^M)\}. \quad (3)$$

2.3 Diversity Learning

After selecting the dominant modality, each modality is encoded by an independent variational encoder. The dominant modality generates feature D^{dm} , and the remaining modalities generate two sets of features, diversity features $\{Z^m\}_{m \neq dm}^M$ and consistent features $\{C^m\}_{m \neq dm}^M$. Then, after two levels of diversity learning, the diversity information of the remaining modalities relative to the dominant modality can be learned while removing redundant information.

Feature-level diversity learning can extract the diversity features between non-dominant and dominant modalities through mutual information and contrastive learning, compress redundant information and separate consistent information from diverse information. Cluster-level diversity learning ensures balanced distribution of each cluster through Uniform Distribution Constraint and mutual information, and enhances the expression of dominant modality from the clustering level. Therefore, the two are synergistically enhanced, enhancing low-level feature expression through feature-level learning, aligning high-level semantic structures through cluster-level learning, and cluster-level can guide feature-level learning through back-propagation to further promote clustering diversity information.

Feature-level Diversity Learning In order to learn the diversity information at the feature level, we propose the following objective function:

$$\min \mathcal{L}_1 = I(X^{dm}; D^{dm}) + \sum_{m \neq dm}^M I(X^m; Z^m) + \sum_{m \neq dm}^M I(Z^m; D^{dm}). \quad (4)$$

Where $I(\cdot)$ is the mutual information measurement. The first two terms describe the compression loss of the raw modalities after passing through the encoders, and the third term aims to learn the diversity information of the dominant modality. Inspired by the Information Bottleneck (IB) theory [24, 25], according to Eq. (4), this method can effectively screen out the minimum sufficient representation that is most critical to the subsequent tasks of each modality under the guidance of the dominant modality, thereby suppressing redundant modal noise and highlighting the diversity information between the modalities.

In order to further effectively strip away the potential diversity information in the non-dominant modality, the dominant modality $\mathbf{D}^{dm}$ is compared and aligned with the consistency features $\{\mathbf{C}^m\}_{m \neq dm}^M$ of the non-dominant modality. Since the encoder network on which the consistency features $\{\mathbf{C}^m\}_{m \neq dm}^M$ and the diversity features $\{\mathbf{Z}^m\}_{m \neq dm}^M$ rely share parameters, the alignment process can more accurately decouple and extract the correct diversity features from the non-dominant modality while maintaining the consistency semantics, thereby improving the expressiveness of the dominant modality. The contrastive loss between the dominant modality $\mathbf{D}^{dm}$ and the consistency features $\mathbf{C}^m$ is calculated as follows:

$$\ell^{dm,m} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(d_i^{dm}, c_i^m)/\tau_1}}{\sum_{s' \in Neg(d_i^{dm}, c_i^m)} e^{s'/\tau_1}}, \quad s_{ij}^{dm,m}(d_i^{dm}, c_j^m) = \frac{(d_i^{dm})^T c_j^m}{\|d_i^{dm}\| \cdot \|c_j^m\|}, \quad (5)$$

where τ_1 is a temperature hyperparameter, $Neg(d_i^{dm}, c_i^m)$ represents the similarity set between negative sample pairs, d_i^{dm} and c_j^m represent the i -th and j -th samples of features $\mathbf{D}^{dm}$ and $\mathbf{C}^m$. Contrastive learning is introduced between the dominant modality and other modalities to facilitate the screening of consistent features and further improve the efficiency of mining diverse information. The common way to calculate contrast loss is to accumulate the contrast loss between all modalities, so the contrast loss function in this section can be expressed as:

$$\mathcal{L}_2 = \frac{1}{2} \left(\sum_{m \neq dm}^M \ell^{dm,m} + \sum_{m \neq dm}^M \ell^{m,dm} \right), \quad (6)$$

to sum up, the loss function of the feature-level diversity learning is expressed as follows:

$$\mathcal{L}_{FDL} = \mathcal{L}_1 + \mathcal{L}_2. \quad (7)$$

Cluster-level Diversity Learning In order to fully capture the semantic coherence between the dominant modality and the non-dominant modality and improve the effectiveness of diversity features in clustering tasks, DDMC further introduces a diversity learning mechanism at the cluster level. Specifically, the diversity features extracted from the non-dominant modality and the dominant modality are input into independent cluster assignment layers to obtain the cluster assignment representation $\{\mathbf{A}^m\}_{m=1}^M$, so as to mine the diversity information between them in the cluster structure, thereby enhancing the clustering expression ability of the dominant modality. To this end, we designed the following objective function:

$$\max \quad \mathcal{L}_3 = \sum_{m \neq dm}^M I(A^m; A^{dm}), \quad (8)$$

at the same time, in order to improve the stability and balance of the clustering results, we further apply the UDC method to adjust the cluster assignment representation $\{\mathbf{A}^m\}_{m=1}^M$ to obtain $\{\mathbf{A}_U^m\}_{m=1}^M$. The objective function of this item is:

$$\min \quad \mathcal{L}_4 = \sum_m^M CE(A^m; A_U^m), \quad (9)$$

where CE is the cross entropy [26]. The core of this method is to carry out standardization processing similar to Sinkhorn-Knopp [27]. This method can convert the soft probability prediction of each sample belonging to each cluster into a more balanced hard label assignment, so that each sample is assigned to only one cluster and the distribution of the number of samples in each cluster is kept as balanced as possible, thereby improving the discriminability and robustness of clustering. The specific implementation details are shown in Appendix A.1. Therefore, the overall objective loss function of cluster-level diversity learning $\mathcal{L}_{CDL}$ is:

$$\mathcal{L}_{CDL} = \mathcal{L}_4 - \mathcal{L}_3. \quad (10)$$

After being processed by the two-levels diversity learning module, the diversity information contained in the remaining modalities can be effectively extracted compared to the dominant modality.

Algorithm 1 :Diversity-oriented Deep Multi-modal Clustering

Input: Multi-modal datasets $\{\mathbf{X}^m\}_{m=1}^M$; Number of clusters K ; Trade-off parameters α and β ; Epoch number E ; Temperature parameters τ_1 .
Initializing the network;
Select the dominant modality by Eq. (3);
for $i = e$ **to** E **do**
 The dominant modality feature $\mathbf{D}^{dm}$, consistency features $\{\mathbf{C}^m\}_{m \neq dm}^M$ and diversity features $\{\mathbf{Z}^m\}_{m \neq dm}^M$ are obtained through modality specific encoders;
 The clustering results $\{\mathbf{A}^m\}_{m=1}^M$ of various modality are obtained through cluster layers;
 Calculate $\mathcal{L}_{FDL}$ with Eq. (29), Eq. (31), Eq. (6) and Eq. (7);
 Calculate $\mathcal{L}_{CDL}$ with Eq. (31), Eq. (9) and Eq. (10);
 Calculate $\mathcal{L}_{DDC}$ by Eq. (12);
 Optimize all parameters by minimizing Eq. (13);
end for
Output:Multi-modal clustering assignment Q .

2.4 Clustering Module

In order to alleviate the information interference and feature confusion problems that may be caused by direct modal fusion, this paper first enhances the dominant modality. The enhanced dominant modality is expressed as:

$$\mathbf{D}_{Enhanced}^{dm} = \text{Concat}(D, Z^1, \dots, Z^m, \dots, Z^M)(m \neq dm). \quad (11)$$

Then, the enhanced dominant mode is clustered by DDC to obtain the cluster assignment matrix Q . DDC [28] is an effective unsupervised clustering method in deep learning, which aims to improve clustering performance by optimizing the dissimilarity measure between sample distributions. It consists of three parts, the first part is intra-cluster compression and inter-cluster separation, the second part is the orthogonality constraint, and the third part is the assignment of simple simplex corners, the loss optimization function of the clustering module is as follows:

$$\mathcal{L}_{DDC} = \frac{1}{K} \sum_{i=1}^{K-1} \sum_{j>i} \frac{\delta_i^T \mathbf{K} \delta_j}{\sqrt{\delta_i^T \mathbf{K} \delta_i \delta_j^T \mathbf{K} \delta_j}} + \text{triu}(Q^T Q) + \frac{1}{K} \sum_{i=1}^{K-1} \sum_{j>i} \frac{\lambda_i^T \mathbf{K} \lambda_j}{\sqrt{\lambda_i^T \mathbf{K} \lambda_i \cdot \lambda_j^T \mathbf{K} \lambda_j}}. \quad (12)$$

Where $\mathbf{K}$ is the Gaussian matrix kernel, δ_i, δ_j denotes the column vectors of the cluster assignment matrix Q , and K is the number of clusters. λ_i is the column vector of $B = [B_{ab}]$, defined as: $B_{ab} = \exp(-\|\alpha_a - e_b\|^2)$.

2.5 Optimization

In summary, DDMC is achieved by minimizing the following loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{DDC} + \alpha \mathcal{L}_{FDL} + \beta \mathcal{L}_{CDL}. \quad (13)$$

Where α, β are trade-off parameters. To optimize the mutual information loss term in Eq. (4), we introduce a variational optimization strategy. Specifically, we estimate the lower bound of the mutual information term $\mathcal{L}_1$ through a variational method and use this as its optimization target, because maximizing its variational lower bound can achieve an unbiased estimate of $\mathcal{L}_1$. The specific implementation process is presented at Appendix A.2 and the training steps of this method are summarized in Algorithm 1.

2.6 Theoretical Analysis

Theorem 1. By optimizing Eq. (7), it is possible to remove the private redundant information of each modality while obtaining the diversity information of the remaining modality relative to the dominant modality.

Proof: For any three modalities ($m \in \{u, dm, v\}$), where dm is the dominant modality. We introduce a simplified assumption: when the training is fully converged, the encoder can accurately capture the joint information. According to the data processing inequality, the mutual information between representations is upper-bounded by that between the original modalities:

$$I(Z^u; D^{dm}) \leq I(X^u; X^{dm}), \quad I(C^u; D^{dm}) \leq I(X^u; X^{dm}), \quad (14)$$

because $Z^u \leftrightarrow X^u \leftrightarrow X^{dm} \leftrightarrow D^{dm}$ forms a Markov chain. Similarly,

$$I(Z^v; D^{dm}) \leq I(X^v; X^{dm}), \quad I(C^v; D^{dm}) \leq I(X^v; X^{dm}). \quad (15)$$

At best, the information that can be shared between each modality and the dominant modality satisfies:

$$\begin{aligned} I(Z^u; D^{dm}) + I(C^u; D^{dm}) &= I(X^u; X^{dm}), \\ I(Z^v; D^{dm}) + I(C^v; D^{dm}) &= I(X^v; X^{dm}). \end{aligned} \quad (16)$$

Since the optimization objective tends to maximize $I(Z^u; D^{dm})$ and $I(C^u; D^{dm})$, while the compression loss enforces the representation to remain concise, the optimal strategy is to assign redundant shared information to a low-dimensional subspace. Therefore, we let the encoding satisfy:

$$I(C^u; D^{dm}) = I(X^u; X^{dm}), \quad I(Z^u; D^{dm}) = 0. \quad (17)$$

That is, C^u carries all common information between X^u and X^{dm} , while Z^u is independent of D^{dm} and contains only modality-specific information. Similarly,

$$I(C^v; D^{dm}) = I(X^v; X^{dm}), \quad I(Z^v; D^{dm}) = 0. \quad (18)$$

At this time, C^u, C^v and D^{dm} share the redundant information between all modalities, while Z^u, Z^v each only carries the remaining independent information. According to the chain rule of mutual information:

$$I(X^u; Z^u, D^{dm}) = I(X^u; D^{dm}) + I(X^u; Z^u | D^{dm}). \quad (19)$$

When $I(Z^u; D) = 0$, it follows that $I(X^u; Z^u | D^{dm}) = I(X^u; Z^u)$. Assuming complete information preservation, we have:

$$I(X^u; Z^u | D^{dm}) = I(X^u; Z^u), \quad I(X^u; Z^u, D^{dm}) = H(X^u), \quad (20)$$

where $H()$ is the entropy. Finally, we obtain:

$$I(Z^u; X^u) = H(X^u) - I(X^u; X^{dm}) = I(X^u; X^u | X^{dm}). \quad (21)$$

That is, Z^u captures the amount of information that X^u still has given X^{dm} , which is exactly the diversity information of X^u relative to X^{dm} . The same is true for Z^v . The above derivation shows that introducing modal contrast loss and mutual information loss under the compression loss constraint of the variational autoencoder will decouple the latent variables of the non-dominant modality from the latent variables of the dominant modality: the shared information is reflected in C^u, C^v , while Z^u, Z^v only contains the diversity information relative to the dominant modality.

3 Experiment

In this segment, we carry out a series of experiments to verify the efficacy of the framework we have proposed. For more experimental data and experimental details, please refer to Appendix A.3.

3.1 Datasets

We evaluate the effectiveness of our proposed method by employing five well-known datasets, including Caltech-3V, Caltech-4V, ESP-Game, Flickr and IAPR. The Caltech image dataset [29], which comprises 1440 samples distributed across 7 distinct classes, is available in three multi-modal variants. **Caltech-3V** incorporates features of Wavelet moments [30], CENSus TRansform hISTogram

Table 2: Clustering results in terms of ACC and NMI on the multi-modal datasets. (The **bold** and underline value are the best and second best result, respectively)

Methods	Caltech-3V		Caltech-4V		ESP-Game		Flickr		IAPR	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KM	46.3	31.3	54.6	46.7	43.2	29.4	40.9	22.5	38.9	17.2
Ncuts(TPAMI'00)	42.6	25.4	67.8	47.6	41.0	25.9	48.4	26.1	41.9	18.9
AmKM	46.9	31.5	44.9	30.6	49.9	34.7	41.0	21.6	40.4	17.0
AmNcuts(TPAMI'00)	43.7	25.5	41.8	24.9	33.5	19.1	48.2	26.2	42.2	18.9
CoregMVSC (NeurIPS'11)	54.4	45.3	64.9	54.5	40.1	28.8	41.0	26.8	35.1	18.4
RMKMC (IJCAI'13)	59.5	49.4	65.5	60.3	44.7	29.7	42.3	23.4	36.4	15.9
SwMC (IJCAI'17)	30.2	23.1	43.7	44.2	43.7	44.2	34.3	34.5	30.2	23.1
ONMSC (AAAI'20)	58.2	56.8	62.3	66.1	17.1	18.1	30.6	16.4	21.6	11.1
SMCMB (TBD'23)	67.2	54.5	74.4	67.0	<u>54.9</u>	<u>40.5</u>	52.8	32.1	34.8	16.4
EAMC (CVPR'20)	38.9	21.4	29.6	16.5	27.1	6.5	30.5	9.1	37.1	16.4
DEMVC (InfoSci'21)	38.7	27.0	48.4	39.7	35.5	21.6	44.8	25.2	30.1	13.8
SiMVC (CVPR'21)	56.9	50.4	61.9	53.6	35.3	16.2	45.6	26.3	42.7	18.5
CoMVC (CVPR'21)	54.1	50.4	56.8	56.8	51.8	38.2	49.3	30.6	46.7	21.5
MFLVC (CVPR'22)	63.1	56.6	73.3	65.2	52.1	39.4	53.8	32.8	<u>47.3</u>	22.6
SEM (NeurIPS'23)	69.2	59.2	82.6	<u>75.3</u>	36.6	23.5	53.1	30.9	42.2	18.9
DIVIDE (AAAI'24)	60.9	53.8	64.3	57.9	46.5	27.0	52.3	<u>33.5</u>	45.6	23.0
SCMVC (TMM'24)	<u>75.9</u>	<u>66.3</u>	<u>84.4</u>	72.9	36.1	24.8	<u>54.2</u>	32.3	46.5	<u>24.1</u>
SSLNMVC (TMM'25)	64.4	58.3	82.1	72.8	44.8	32.3	51.2	33.0	46.4	24.0
DDMC	76.7	68.8	90.3	82.7	60.9	40.9	58.7	36.5	49.5	28.3
Ours vs BestCompared	0.8↑	2.5↑	5.9↑	7.4↑	6.0↑	0.4↑	4.5↑	3.0↑	2.2↑	4.2↑

(CENTRIST) [31] and Local Binary Pattern [32], treating each feature type as a separate modality. **Caltech-4V** builds upon Caltech-3V by introducing an additional feature, namely Generalized Search Trees [33]. **ESP-Game** [34] dataset is derived from an online image tagging game, containing 11,032 images across 7 categories, with each sample having three modality descriptions. **Flickr** [22] dataset is a widely used multi-modal dataset for image retrieval, containing 12154 sample across 6 categories. It utilizes the same three modalities as ESP-Game. **IAPR** [35] is a comprehensive multi-modal image dataset, containing 6 categories with 7,855 samples and two features.

Table 1: Details about the Multi-modal Datasets

Dataset	Samples	Clusters	Dimension
Caltech-3V	1440	7	40/254/928
Caltech-4V	1440	7	40/254/928/512
ESP-Game	11032	7	300/300/300
Flickr	12154	6	100/100/100
IAPR	7855	6	100/100

3.2 Compared Methods

To further substantiate the merits of our proposed approach, we conducted comparative experiments with a comprehensive set of eighteen baseline methods, classified into three types: **Single-Modal Clustering Methods**: K-Means(KM), Normalized Cuts(Ncuts) [36], All-modalities KMeans(AvKM), All-modal Normalized Cuts(AvNcuts); **Traditional Multi-modal Clustering Methods**: CoregMVSC [37], RMKMC [38], SwMC [39], ONMSC [40], SMCMB [41]; **Deep Multi-modal Clustering Methods**: MVSCN [42], DEMVC [43], SiMVC [44], CoMVC [44], MFLVC [45], SEM [20], DIVIDE [46], SCMVC [47], SSLNMVC [48]. All comparison method codes are obtained from the original authors through their official GitHub repositories or personal homepages.

Evaluation Metrics The clustering performance is meticulously assessed by employing two extensively utilized metrics: Accuracy (ACC) and Normalized Mutual Information (NMI) [12]. [49] The higher the value of these two indicators, the better the performance of our model in terms of clustering accuracy and consistency.

3.3 Clustering Performance Analysis

Table 2 demonstrates the clustering effectiveness of our proposed method on five publicly available datasets, with the clustering outcomes in terms of ACC and NMI being reported. It is observed that

Table 3: Ablation Study on different multi-modal datasets (The **bold** and underline value are the best and second best result, respectively).

Methods	Caltech-3V		Caltech-4V		ESP-Game		Flickr		IAPR	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
(1) $\mathcal{L}_{DDC}$	70.2	53.2	77.8	69.6	35.3	14.4	49.0	30.7	39.1	18.0
(2) $\mathcal{L}_{DDC} + \mathcal{L}_{FDL}$	71.4	61.3	79.4	73.5	<u>54.9</u>	<u>35.5</u>	55.2	35.4	<u>48.8</u>	<u>28.1</u>
(3) $\mathcal{L}_{DDC} + \mathcal{L}_{CDL}$	<u>75.4</u>	<u>64.3</u>	<u>82.1</u>	<u>81.8</u>	36.6	22.0	<u>55.3</u>	<u>36.3</u>	45.4	24.0
(4) All Modules (The Proposed Method)	76.7	68.8	90.3	82.7	60.9	40.9	58.7	36.5	49.5	28.3

our method significantly outperforms the compared single/all-view, traditional and deep multi-modal clustering methods. The remarkable results demonstrate the strong clustering capability of our method, which is attributed to the two levels diversity learning that is capable of acquiring diverse knowledge and eliminating redundant information.

The proposed method in this study demonstrates remarkable adaptability. On the Caltech dataset, as the number of modalities increases, the ACC and NMI of the model show a steady upward trend, owing to the rich complementary information contained in multi-modal data. As the number of modalities increases, the model can obtain data features from more dimensions. This indicates that the proposed method is capable of efficiently integrating information from different modalities and fully leveraging the unique features and clustering information contained in each modality, thereby significantly enhancing the accuracy and consistency of clustering. However, such an expected improvement in clustering performance is conspicuously absent when observing the outcomes of the EAMC and DEMC methods. These findings indicate that these methods might not be sufficiently robust when confronted with variations in the number of modalities. Compared with the best competitors, our method also obtains excellent performance. For example, on the ESP-Game dataset, our method demonstrates a substantial improvement in both ACC and NMI compared to the second-best method (SMCMB), achieving increases of 6.0% and 0.5% respectively. In terms of ACC and NMI results on the five tested multi-modality datasets, the proposed method achieves average improvements of 3.88% and 3.52% over the second-best comparison method, respectively. This highlights the stability of our method across different datasets.

3.4 Ablation Study

In this part, we give an ablation study to further show the effectiveness of different components of our method. The DDMC includes a clustering module and two diversity learning modules, which means there are four possible combinations. The outcomes of these combinations are presented in Table 3. The results indicate that the performance generally improves as more modules are incorporated into the model. For instance, on the Caltech-3V dataset, the baseline model achieved an ACC of 70.2% and a NMI of 53.2%. In contrast, the model that includes all modules saw a significant enhancement in performance, with an ACC of 76.7% and a NMI of 68.8%. Similar trends are observed across other datasets, with particularly notable performance enhancements in the Caltech-4V and IAPR datasets when all modules are integrated into the model. Specifically, the model equipped with all modules achieved an ACC of 90.3% and a NMI of 82.7% on the Caltech-4V dataset, and an ACC of 49.5% and a NMI of 28.3% on the IAPR dataset. These results underscore the superior performance of the model when all components are utilized.

3.5 Enhancement *vs* Fusion

In addition, we replaced the final enhancement step with the commonly used modality fusion strategy, and the comparison results are shown in Figure 4. Experiments show that the enhancement strategy outperforms the fusion method on all datasets; especially on the ESP-Game dataset, the ACC is improved by about 11%. This result shows that compared with modality fusion, dominant modality enhancement can significantly improve clustering performance, while the noise introduced in the fusion process will weaken the role of the dominant modality and reduce the clustering effect, further verifying the effectiveness of the DDMC.

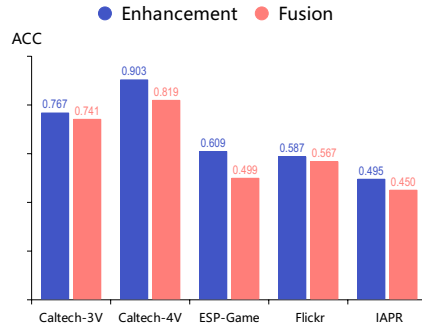


Figure 4: Enhancement *vs* Fusion

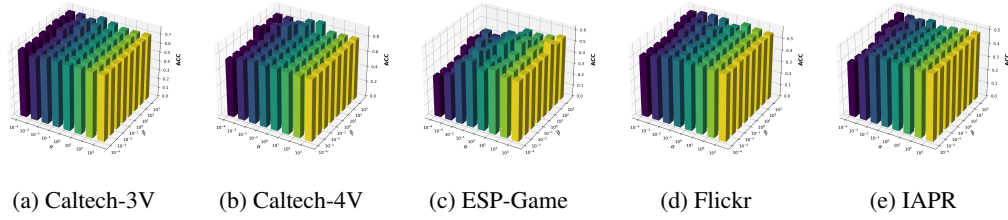


Figure 5: Parameter analysis of DDMC on different datasets.

3.6 Parameter Sensitivity

In the proposed method, we use two trade-off parameters α , β to balance $\mathcal{L}_{FDL}$, $\mathcal{L}_{CDL}$. To thoroughly investigate the sensitivity of these parameters, we conducted extensive experiments on all datasets with various parameter configurations. Specifically, we utilized a grid search strategy to optimize the values of α and β within the range $[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3]$. The clustering ACC results are illustrated in Figure 5. As can be seen, our method exhibits robust performance across all datasets, with minimal performance degradation under most parameter settings. This suggests that our method is relatively insensitive to parameter fluctuations and maintains stable performance overall. Therefore, DDMC can have good performance in a large range of parameter values.

4 Conclusion and Limitations

This paper introduces a novel deep multi-modal clustering framework that leverages a dominant-modality enhancement strategy to mitigate noise from conventional feature-fusion. Rather than fusing all modalities indiscriminately, we identify the highest-quality modality as dominant, perform two-level diversity learning to extract diversity information from the remaining modalities, and augment the dominant modality accordingly. This strategy significantly improves the clustering performance, especially when the multi-modal data is unevenly distributed or has large quality differences.

In addition, our method is not without limitations. Although SI is an excellent indicator, there may still exist other theoretically feasible indicators. We will continue to study them in the future and introduce more robust dominant modality selection mechanisms. We also realize that this method performs poorly when dealing with incomplete modal data, especially when some modalities are missing or the data is unbalanced. In future research, We will explore more powerful selection strategies and effective strategies for extracting diversity information, such as combining multiple indicators or adaptive optimization methods, to ensure that the selection of the dominant modality is more convincing.

Acknowledgment

The authors thank anonymous reviewers for their constructive comments. This work was supported by Henan Province Outstanding Youth Science Fund Program under Grant 252300421223, National Natural Science Foundation of China under Grant 62206254, 62325602, 62036010, and China Postdoctoral Science Foundation under Grant 2024T170843 and 2023M743186.

References

- [1] A. Roy, P. Rivas, and M. N. Tayba, “Unsupervised machine learning methods for diagnosing autism spectrum disorder using multimodal data: A survey,” in *Proc. Int. Conf. Comput. Sci. Comput. Intell.*, 2022, pp. 41–52.
- [2] D. Townsell, T. Banerjee, L. Chen, and M. Raymer, “Advancing chest x-ray diagnostics via multi-modal neural networks with attention,” in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2024, pp. 1–4.
- [3] A. Kumar, “Deep learning for multi-modal medical imaging fusion: Enhancing diagnostic accuracy in complex disease detection,” *Int J Eng Technol Res Manag*, vol. 6, no. 11, p. 183, 2022.

- [4] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, 2021.
- [5] X. Zhang, Y. Gong, J. Lu, J. Wu, Z. Li, D. Jin, and J. Li, "Multi-modal fusion technology based on vehicle information: A survey," *IEEE Trans. Intell. Veh.*, vol. 8, no. 6, pp. 3605–3619, 2023.
- [6] Z. Huang, X. Xu, J. Ni, H. Zhu, and C. Wang, "Multimodal representation learning for recommendation in internet of things," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 675–10 685, 2019.
- [7] Y. Tong, "Research on deep clustering and multimodal deep learning service recommendation system for large-scale user data," in *Int. Conf. Integr. Circuits Commun. Syst.*, 2025, pp. 1–5.
- [8] J. Li, C. Yang, G. Ye, and Q. V. H. Nguyen, "Graph neural networks with deep mutual learning for designing multi-modal recommendation systems," *Inf. Sci.*, vol. 654, p. 119815, 2024.
- [9] A. Kumar and H. D. III, "A co-training approach for multi-view spectral clustering," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 393–400.
- [10] R. Zhou and Y. Shen, "End-to-end adversarial-attention network for multi-modal clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14 607–14 616.
- [11] Y. Li and H. Liao, "Multi-view clustering via adversarial view embedding and adaptive view fusion," *Appl. Intell.*, vol. 51, pp. 1201–1212, 2021.
- [12] S. Hu, G. Zou, C. Zhang, Z. Lou, R. Geng, and Y. Ye, "Joint contrastive triple-learning for deep multi-view clustering," *Inf. Process. Manag.*, vol. 60, p. 103284, 2023.
- [13] G. Ke, G. Chao, X. Wang, C. Xu, Y. Zhu, and Y. Yu, "A clustering-guided contrastive fusion for multi-view representation learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, pp. 2056–2069, 2024.
- [14] H. Meng, D. Huang, H. Wang, H. Yang, M. Al-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proc. Int. Workshop Audio/Vis. Emot. Chall.*, 2013, pp. 21–30.
- [15] A. Samareh, Y. Jin, Z. Wang, X. Chang, and S. Huang, "Detect depression from communication: How computer vision, signal processing, and sentiment analysis join forces," *IJSE Trans. Healthc. Syst. Eng.*, vol. 8, pp. 1–42, 2018.
- [16] X. Liu, L. Liu, Q. Liao, S. Wang, Y. Zhang, W. Tu, C. Tang, J. Liu, and E. Zhu, "One pass late fusion multi-view clustering," in *Proc. Int. Conf. Mach. Learn.*, 2021.
- [17] M. Morales, S. Scherer, and R. Levitan, "A linguistically-informed fusion approach for multimodal depression detection," in *Proc. Workshop Comput. Linguist. Clin. Psychol.: From Keyboard Clin.*, 2018, pp. 13–24.
- [18] S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear, "Cross-cultural detection of depression from nonverbal behaviour," in *Proc. IEEE Int. Conf. Auto. Face Gesture Recognit.*, vol. 1, 2015, pp. 1–8.
- [19] H. Shalu, H. P. H. S. CN, A. Das, S. Majumder, A. Datar, S. M. MS, A. Das, and J. Kadiwala, "Depression status estimation by deep learning based hybrid multi-modal fusion model," *CoRR*, vol. abs/2011.14966, 2020.
- [20] J. Xu, S. Chen, Y. Ren, X. Shi, H. Shen, G. Niu, and X. Zhu, "Self-weighted contrastive learning among multiple views for mitigating representation degeneration," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023.
- [21] D. Qiao, G. Liu, T. Lv, W. Li, and J. Zhang, "Marine vision-based situational awareness using discriminative deep learning: A survey," *J. Mar. Sci. Eng.*, vol. 9, no. 4, 2021.
- [22] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proc. Int. ACM Conf. Multimedia Inf. Retr. MIR.*, 2008, pp. 39–43.
- [23] J. A. H. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *J. R. Stat. Soc.*, vol. 28, no. 1, pp. 100–108, 1979.
- [24] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Annu. Allerton Conf. Commun.*, 1999, pp. 368–377.

- [25] S. Hu, Z. Lou, X. Yan, and Y. Ye, “A survey on information bottleneck,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5325–5344, 2024.
- [26] A. Mao, M. Mohri, and Y. Zhong, “Cross-entropy loss functions: theoretical analysis and applications,” in *Proc. Int. Conf. Mach. Learn.*, 2023.
- [27] P. A. Knight, “The sinkhorn-knopp algorithm: Convergence and applications,” *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 1, pp. 261–275, 2008.
- [28] M. Kampffmeyer, S. Løkse, F. M. Bianchi, L. Livi, A. Salberg, and R. Jenssen, “Deep divergence-based approach to clustering,” *Neural Netw.*, vol. 113, pp. 91–101, 2019.
- [29] L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2004, pp. 178–178.
- [30] D. Shen and H. H. Ip, “Discriminative wavelet shape descriptors for recognition of 2-d patterns,” *Pattern Recognit.*, vol. 32, no. 2, pp. 151–165, 1999.
- [31] J. Wu and J. M. Rehg, “Centrist: A visual descriptor for scene categorization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, 2011.
- [32] L. Wolf, T. Hassner, and Y. Taigman, “Descriptor based methods in the wild,” *Proc. Eur. Conf. Comput. Vis.*, 10 2008.
- [33] A. Friedman, “Framing pictures: The role of knowledge in automatized encoding and memory for gist,” *J. Exp. Psychol.-Gen.*, vol. 108, pp. 316–55, 09 1979.
- [34] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, “Nus-wide: a real-world web image database from national university of singapore,” in *Proc. ACM Int. Conf. Image Video Retr.*, 2009.
- [35] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, “The iapr tc12 benchmark: A new evaluation resource for visual information systems,” *Jt. Int. Conf. Comput. Linguist.*, 10 2006.
- [36] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.
- [37] A. Kumar, P. Rai, and H. Daume, “Co-regularized multi-view spectral clustering,” in *Adv. Neural Inf. Process. Syst.*, vol. 24, 2011.
- [38] X. Cai, F. Nie, and H. Huang, “Multi-view k-means clustering on big data,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2598–2604.
- [39] F. Nie, J. Li, and X. Li, “Self-weighted multiview clustering with multiple graphs,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 2564–2570.
- [40] S. Zhou, X. Liu, J. Liu, X. Guo, Y. Zhao, E. Zhu, Y. Zhai, J. Yin, and W. Gao, “Multi-view spectral clustering with optimal neighborhood laplacian matrix,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6965–6972.
- [41] J. Lao, D. Huang, C.-D. Wang, and J.-H. Lai, “Towards scalable multi-view clustering via joint learning of many bipartite graphs,” *IEEE Trans. Big Data.*, vol. 10, no. 1, pp. 77–91, 2024.
- [42] Z. Huang, J. T. Zhou, X. Peng, C. Zhang, H. Zhu, and J. Lv, “Multi-view spectral clustering network,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 2563–2569.
- [43] J. Xu, Y. Ren, G. Li, L. Pan, C. Zhu, and Z. Xu, “Deep embedded multi-view clustering with collaborative training,” *Inf. Sci.*, vol. 573, pp. 279–290, 2021.
- [44] D. J. Trosten, S. Løkse, R. Jenssen, and M. Kampffmeyer, “Reconsidering representation alignment for multi-view clustering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1255–1265.
- [45] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, and L. He, “Multi-level feature learning for contrastive multi-view clustering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16 030–16 039.
- [46] Y. Lu, Y. Lin, M. Yang, D. Peng, P. Hu, and X. Peng, “Decoupled contrastive multi-view clustering with high-order random walks,” in *Proc. AAAI Conf. Artif. Intell.*, 2024.
- [47] S. Wu, Y. Zheng, Y. Ren, J. He, X. Pu, S. Huang, Z. Hao, and L. He, “Self-weighted contrastive fusion for deep multi-view clustering,” *IEEE Trans. Multimedia.*, vol. 26, pp. 9150–9162, 2024.

- [48] W. Yan, T. Yang, and C. Tang, “Self-supervised semantic soft label learning network for deep multi-view clustering,” *IEEE Trans. Multimedia.*, pp. 1–13, 2025.
- [49] Y. Gan, Y. You, J. Huang, S. Xiang, C. Tang, W. Hu, and S. An, “Multi-view clustering via multi-stage fusion,” *IEEE Trans. Multimed.*, pp. 1–13, 2025.
- [50] R. Jonker and T. Volgenant, “Improving the hungarian assignment algorithm,” *Oper. Res. Lett.*, vol. 5, no. 4, pp. 171–175, 1986.
- [51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

A Appendix

A.1 The Details of Uniform Distribution Constraint

The role of UDC is to force the edge distribution of each cluster to be as close to uniform as possible when generating hard labels from the soft clustering distribution. Firstly, perform exponentiation operations on the original cluster distribution matrix $\{\mathbf{A}^m\}_{m=1}^M$ to enhance the diagonal weights. Then, define two edge vectors $r \in \mathbb{R}^K$ and $c \in \mathbb{R}^N$. The goal is to make $P = \text{diag}(r)A \text{diag}(c)$ satisfy the following requirements:

$$P\mathbf{1}_K = \frac{1}{N}\mathbf{1}_N, \quad P^\top \mathbf{1}_N = \frac{1}{K}\mathbf{1}_K, \quad (22)$$

That is, P approximates a ‘coupling matrix’ that is uniform for both rows and columns. Then iterative calculation is carried out through the Sinkhorn-Knopp algorithm:

$$r^{(t+1)} = \frac{\frac{1}{K}\mathbf{1}_K}{A^{(t)}}, \quad c^{(t+1)} = \frac{\frac{1}{N}\mathbf{1}_N}{A^\top r^{(t+1)}} \quad (23)$$

Until convergence, transform the normalized matrix P back to the $N \times K$ shape and take the maximum index of the n -th row to obtain the discrete label: $A_{Un} = \arg \max_k P_{n,k}$. The above process not only ensures the discreteness of the output labels, but also approximately satisfies the uniform distribution constraint of each cluster globally.

A.2 Optimization

Taking a modality X^1 as an example, based on the definition of mutual information, we can get the following formula:

$$I(X^1; Z^1) = \int_{z^1, x^1} p(z^1, x^1) \log \frac{p(z^1, x^1)}{p(z^1)p(x^1)} = \int_{z^1, x^1} p(z^1, x^1) \log \frac{p(z^1 | x^1)}{p(z^1)}. \quad (24)$$

Since the posterior distribution $p(z^1, x^1)$ cannot be solved directly, another scalable distribution $q(z^1)$ is used to approximate $p(z^1)$. By minimizing the Kullback-Leibler (KL) divergence between the two distributions, $q(z^1)$ is optimized step by step to make it similar to $p(z^1)$. Specifically, since KL is non-negative, we can get:

$$\begin{aligned} \text{KL}[p(z^1), q(z^1)] &= \int_{z^1} p(z^1) \log \frac{p(z^1)}{q(z^1)} > 0 \\ \implies \int_{z^1} p(z^1) \log p(z^1) &> \int_{z^1} p(z^1) \log q(z^1) \end{aligned} \quad (25)$$

Now, Eq.(24) can be estimated as:

$$I(X^1; Z^1) < \int_{z^1, x^1} p(z^1, x^1) \log \frac{p(z^1 | x^1)}{q(z^1)} < \int_{z^1, x^1} p(z^1 | x^1) p(x^1) \log \frac{p(z^1 | x^1)}{q(z^1)} \quad (26)$$

Since $\{\mathbf{Z}^m\}_{m \neq dm}^M$ does not contain the dm -th term and the dominant modality is not distinguished in the calculation of compression loss, for the sake of simplicity, we temporarily let $Z^{dm} = D^{dm}$, so that the compression loss in Eq.(4) can be estimated as:

$$\sum_{m \neq dm}^M I(X^m; Z^m) + I(X^{dm}; D) < \sum_m^M \int_{z^m, x^m} p(z^m | x^m) \log \frac{p(z^m | x^m)}{q(z^m)} \quad (27)$$

In order to remove unnecessary items, Monte Carlo sampling is used to replace $\sum_m^M p(x^m)$. Therefore, Eq.(27) can be further expressed as:

$$\sum_{m \neq dm}^M I(X^m; Z^m) + I(X^{dm}; D) < \frac{1}{N} \sum_i^N \left\{ \sum_m^M \int_{z^m} p(z^m | x^m) \log \frac{p(z^m | x^m)}{q(z^m)} \right\} \quad (28)$$

where N is the number of data samples. Assuming that $\sum_m^M p(z^m|x^m)$ conforms to the Gaussian distribution, the mean μ and variance σ of each modality can be learned through the encoder. For the convenience of calculation, we reparameterize $\sum_m^M z^m$ by $\sum_m^M z^m = \mu(x^m) + \sigma(x^m) * \epsilon$, where ϵ is the standard normal distribution. Therefore, the objective loss function for calculating the compression loss in Eq.(4) can be expressed as:

$$\begin{aligned} \sum_{m \neq dm}^M I(X^m; Z^m) + I(X^{dm}; D) &< \frac{1}{N} \sum_i^N \left\{ \sum_m^M \mathbb{E}_\epsilon \log \frac{p(z^m|x^m)}{q(z^m)} \right\} \\ &\approx \frac{1}{N} \sum_i^N \mathbb{E}_\epsilon \left\{ \sum_m^M KL[p(z^m|z^m), q(x^m)] \right\} \end{aligned} \quad (29)$$

where $\mathbb{E}_\epsilon$ represents mathematical expectation. For the calculation of diversity information loss in Eq.(4), we use the discrete joint probability estimation for calculation:

$$\sum_{m \neq dm}^M I(Z^m; D^{dm}) = \sum_{m \neq dm}^M \sum_{i=1}^{d_m} \sum_{j=1}^{d_m} p^{m,dm}(i,j) \log \left(\frac{p^{m,dm}(i,j)}{p^m(i) p^{dm}(j)} \right) \quad (30)$$

where:

$$\begin{aligned} p^m(i) &= \sum_{j=1}^{d_m} p^{m,dm}(i,j), p^{dm}(j) = \sum_{i=1}^{d_m} p^{m,dm}(i,j), \\ p^{m,dm}(i,j) &= \frac{1}{2N} \sum_{n=1}^N [Z_n^m(i) \cdot D_n^{dm}(j) + D_n^{dm}(i) \cdot Z_n^m(j)]. \end{aligned} \quad (31)$$

Similarly, the calculation of cluster assignment mutual information $\mathcal{L}_3$ can also be optimized by Eq.(31).

A.3 Further Analysis

Evaluation Metrics ACC is used to evaluate the degree of correspondence between the assigned cluster labels and the true labels of the data points, essentially measuring the extent to which the clustering algorithm correctly identifies the inherent groupings within the dataset, which is defined as:

$$ACC = \frac{\sum_{i=1}^n \delta(p_i, \text{map}(q_i))}{n} \quad (32)$$

where n denotes the number of samples, $\delta(i,j) = 1$ if $i = j$ ($\delta(i,j) = 0$, otherwise), and $\text{map}(q_i)$ represents the clustering result q_i being matched to the ground truth p_i through Hungarian algorithm[50]. NMI quantifies the mutual dependence between the clustering results and the ground truth, taking into account the distribution of data points across different clusters, formulated by:

$$NMI(\Omega, C) = \frac{I(\Omega; C)}{(H(\Omega) + H(C))/2} \quad (33)$$

where Ω indicates the original ground truth, and C denotes the clustering information, $H()$ represents the entropy of the clustering result, and $I(;)$ indicates mutual information between clustering results.

Details of the Compared Methods **K-Menas:** An unsupervised learning algorithm that divides data into K clusters so that the similarity of data points within clusters is high and the similarity between clusters is low.

All-modal K-Means: This algorithm concatenate all modal features and then executes KMenas algorithm

Normalied Cuts: [36] An image segmentation algorithm based on graph theory is proposed to minimize the normalized cut value to optimize the image region partitioning, while considering the inter-group dissimilarity and intra-group similarity.

All-modal Normalized Cuts: The method uses Normalized Cuts for the joint representation of each data point to achieve clustering.

CoregMVSC: [37] A co-regularized multi-view spectral clustering method that enhance accuracy by enforcing clustering consistency across views.

RMKMC: [38] A multi-view multiple kernel clustering framework based on the restarted strategy and self-guiding mechanism.

SwMC: [39] A self-weighted multi-view clustering algorithm that iteratively optimizes the target similarity matrix and view weights to effectively integrate multi-view data for clustering without requiring additional parameters.

ONMMSC: [40] A multi-view spectral clustering algorithm searches for the optimal matrix within the neighborhood of the linear combination of Laplacian matrices, breaking through the limitations of traditional methods to enhance clustering performance.

SMCMB: [41] This method mining rich information in multi-view data by joint learning of multiple bipartite graphs, and maintaining high efficiency on large-scale data sets, the time and space complexity is close to linear.

EAMC: [10] An end-to-end adversarial attention network that aligns potential feature distributions and quantifies modal importance, respectively, through adversarial learning and attention mechanisms.

DEMVC: [43] A multi-view clustering algorithm that utilizes common and complementary information from multiple views to achieve better clustering performance through deep embedded representation learning and collaborative training mechanisms.

SiMVC: [44] A deep multi-view clustering baseline model that does not require alignment of representation distributions

CoMVC: [44] A deep multi-view clustering algorithm that selectively aligns representations of different views at the sample level by contrasting learning frameworks.

MFLVC: [45] A multi-level feature learning framework is proposed, which can effectively reduce the interference of view private information by separating reconstruction target and consistency target.

DIVIDE: [46] A novel robust multi-view clustering method, which identifies global data pairs via high-order random walks and employs a decoupled contrastive learning framework to perform intra-view and inter-view contrastive learning in separate embedding spaces, thereby enhancing clustering performance and robustness against missing views.

SSLNMVC: [48] A deep multi-view clustering method that enhances the consistency of multi-view features through a consensus high-level feature learning module and aligns view-specific and view-consensus semantic labels using a self-supervised semantic calibration module.

SEM: [20] This method solves the representation degradation problem caused by contrast learning in multi-view scenarios through self-weighting and information reconstruction strategies.

SCMVC: [47] The method establishes a hierarchical feature fusion framework and a self-weighted contrastive fusion approach, effectively separating the consistency objective from the reconstruction objective.

Implementation Details Our experiments were conducted on a Windows 10 operating system, utilizing a powerful configuration equipped with 96 GB of system memory and a high-performance NVIDIA GeForce RTX 4090D GPU. We implemented the proposed framework using the PyTorch platform[51]. For all datasets, the training batch size was uniformly set to 512, and we utilized the Adam optimization algorithm with an initial learning rate of 0.0003. The configuration of parameters in the proposed model is detailed as follows. The hyperparameters α and β are tuned to values ranging from 0.0001 to 1000, with each value being a power of 10. Given that 100 epochs proved to be ample for the training convergence of algorithm, we accordingly trained the model from the beginning up to 100 epochs. To enhance robustness and circumvent local minima, we trained the proposed model 10 times, reporting the clustering outcome with the minimal clustering loss. For all datasets, we utilized modal-specific variational encoders comprising three fully connected layers, each layer consists of a batch normalization layer and a RELU layer. The second layer and the output layer are set to 512. The clustering layer adopts a fully connected layer with a softmax layer. The parameter of the dropout layer is set to 0.1. The temperature hyperparameter in the comparative learning is set to 0.5.

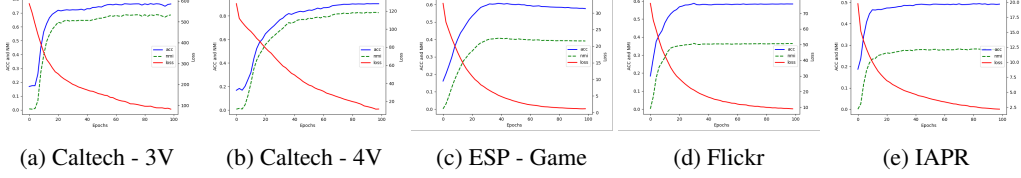


Figure 6: Convergence analysis of DDMC on the datasets in the order of Caltech - 3V Caltech - 4V, ESP - Game, Flickr and IAPR, respectively.

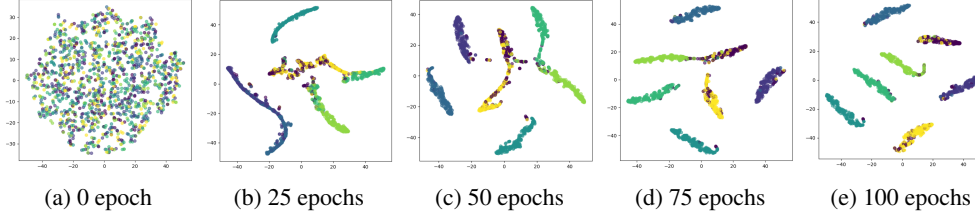


Figure 7: Evolution of cluster assignments during training on the Caltech-4V dataset.

Convergence Analysis To evaluate the convergence properties of our proposed approach, we conducted a comprehensive series of experiments across a diverse range of datasets, including Caltech-3V, Caltech-4V, ESP-Game, Flickr, and IAPR datasets. As depicted in Figure 6, we present the objective loss values in conjunction with the clustering performance metrics, specifically ACC and NMI, throughout the training epochs. It is evident that the loss values experience a pronounced decline during the initial stages of training, particularly within the first 25 epochs. This rapid decrease indicates that the model is effectively learning and adjusting its parameters to minimize the loss function. Subsequently, the loss values begin to stabilize, suggesting that the model has reached a point of diminishing returns in terms of further loss reduction. The results consistently demonstrate that convergence for all metrics is achieved after the 100-th epoch threshold. This uniform convergence pattern across the diverse datasets highlights the robustness of our algorithmic iterations in attaining a balanced state, thereby ensuring consistent performance metrics irrespective of dataset variations.

T-SNE Visualization Analysis In order to further substantiate the efficacy of our proposed framework, we employed the widely recognized t-sne tool to visualize the clustering outcomes of DDMC on the Caltech-4V dataset. The joint training of our approach yields satisfactory clustering results. To provide an intuitive depiction of the training evolution, we carried out t-sne visualizations of the clustering results at different training epochs, namely the 0-th, 25-th, 50-th, and 100-th epoch, as illustrated in Figure 7. In this figure, distinct colors are assigned to represent the various clusters generated by the DDC module. It is evident that as the number of training epochs increases, the clustering assignments tend to become more compact and well-separated from each other. This observation indicates the effectiveness of our proposed approach in enhancing clustering performance throughout the training process.

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 sections accurately reflect the contribution of the paper. For details, please refer to the abstract and Section 1.

Guidelines:

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

2. Limitations

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

Answer: [\[Yes\]](#)

Justification: The limitations of the paper are explained in Section 4. Please see Section 4 for details.

Guidelines:

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

3. Theory assumptions and proofs

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

Answer: [\[Yes\]](#)

Justification: The paper provides a detailed description of the proposed theory and optimization. For details, please see Section 2.6 and Appendix A.1.

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [\[Yes\]](#)

Justification: The experiments in Section 3 are all real and effective. Please see Section 3 for details.

Guidelines:

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

5. Open access to data and code

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

Answer: [Yes]

Justification: Data and code are publicly available.

Guidelines:

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

6. Experimental setting/details

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

Answer: [Yes]

Justification: Please see Section 3 and Appendix A.3 for details.

Guidelines:

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

7. Experiment statistical significance

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

Answer: [No]

Justification: The paper does not report error bars.

Guidelines:

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

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

8. Experiments compute resources

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

Answer: [Yes]

Justification: Please see Appendix A.3 for details.

Guidelines:

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

9. Code of ethics

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

Answer: [Yes] ,

Justification: This article complies with the NeurIPS ethical standards.

Guidelines:

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

10. Broader impacts

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

Answer: [NA]

Justification: The research work in this paper will not have any impact on society.

Guidelines:

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

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

11. Safeguards

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

Answer: [NA]

Justification: There is no such risk in the paper.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [NA]

Justification: This paper does not use existing assets.

Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

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

Answer: [NA]

Justification: This paper does not release new assets.

Guidelines:

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

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

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

Answer: [NA]

Justification: The paper involved no crowdsourcing and no research with human subjects.

Guidelines:

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

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

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

Answer: [NA]

Justification: The paper involved no crowdsourcing and no research with human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: The core methodological development of this study did not involve LLM as any significant, original, or nonstandard component.

Guidelines:

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