Empowering Self-Balance of Deep Information Bottleneck for Multimodal Clustering

Youwei Wang[1] , **Qiwei Miao**[1] , **Haichuan Fang**[2] , **Yangdong Ye**[2] and **Shizhe Hu** [2*]
[1]School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou, China
[2]School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China
youweiwang@zzu.edu.cn, ieshizhehu@gmail.com

Abstract

Multimodal clustering (MMC) focuses on learning consistent representations through fusing discriminative features from each modality in an unsupervised fashion. Recently, information bottleneck-based MMC methods transform the representation learning into a non-redundant multimodal feature puzzle process. However, they depend on manually setting the trade-off parameter β to compress task-irrelevant features while preserving task-relevant features, which severely impairs final clustering results. In addition, manually setting β overlooks both the *information quality balance* and the *information quantity balance*, hindering the learning of compact and meaningful representations for discovering cluster patterns. In this work, we propose a novel β_{qq}-guided information bottleneck (β_{qq} IB) for addressing the aforementioned problems. The core of β_{qq} IB is a self-balance mechanism that consists of a $\beta_{quality}$ component and a $\beta_{quantity}$ component. The $\beta_{quality}$ component aims to achieve an information quality balance by modeling the relation between task characteristics and data sample scale. Meanwhile, the $\beta_{quantity}$ component intends to realize information quantity balance through modeling the dynamic changes of data scale within the deep variational architect. Experiments on six benchmark datasets demonstrate the superiority of β_{qq} IB method. To the best of our knowledge, this is the first work to investigate self-balance learning in IB-based multimodal clustering.

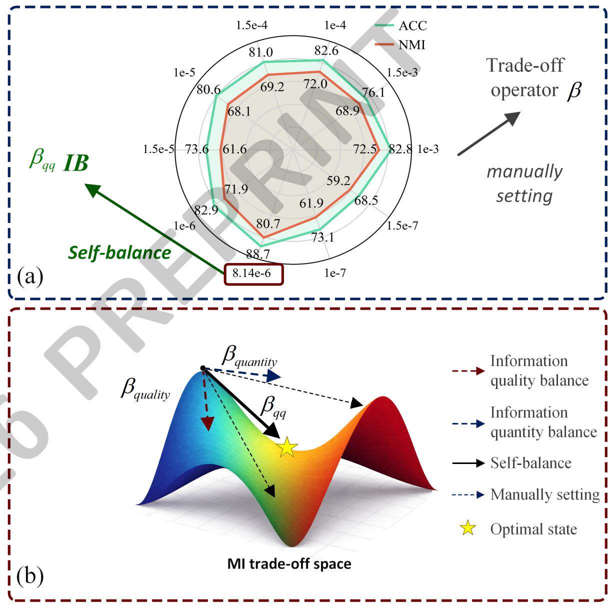

Figure 1: The intractable issues of manually setting trade-off operator β in IB-based MMC methods. (a) Clustering results comparison on Caltech_4V using the same IB-based method with automatically learned β versus manually tuned β. (b) Overlook information quality balance and information quantity balance leads to suboptimal solutions in the mutual information (MI) minimum-maximum trade-off space.

1 Introduction

Multimodal clustering (MMC) is a hot topic in the era of large language models, and it focuses on learning semantically consistent representations by fusing the most discriminative features from each modality to identify cluster patterns [Chowdhury *et al.*, 2025; Fang *et al.*, 2023]. MMC has been successfully used into many real-world application scenarios, such as autonomous drone operation and cross-camera object identification. Of all the MMC methods, information bottleneck (IB)-based MMC is one of the effective methods working on by exploring information compression-preservation mechanism to learn compact and meaningful representation from different modalities.

Existing IB-based MMC methods can be broadly categorized into two groups: (1) Classical IB method: Employing classical machine learning methods for modality feature mining. For instance, a Markov process is used to compute within-modality data similarity [Tishby and Slonim, 2000]. Afterwards, an information-theoretic function is employed to capture low-level features within each modality and high-level features across modalities [Yan *et al.*, 2020]. A weight learning-based MMC framework has been proposed to effec-

*Correspongding Author

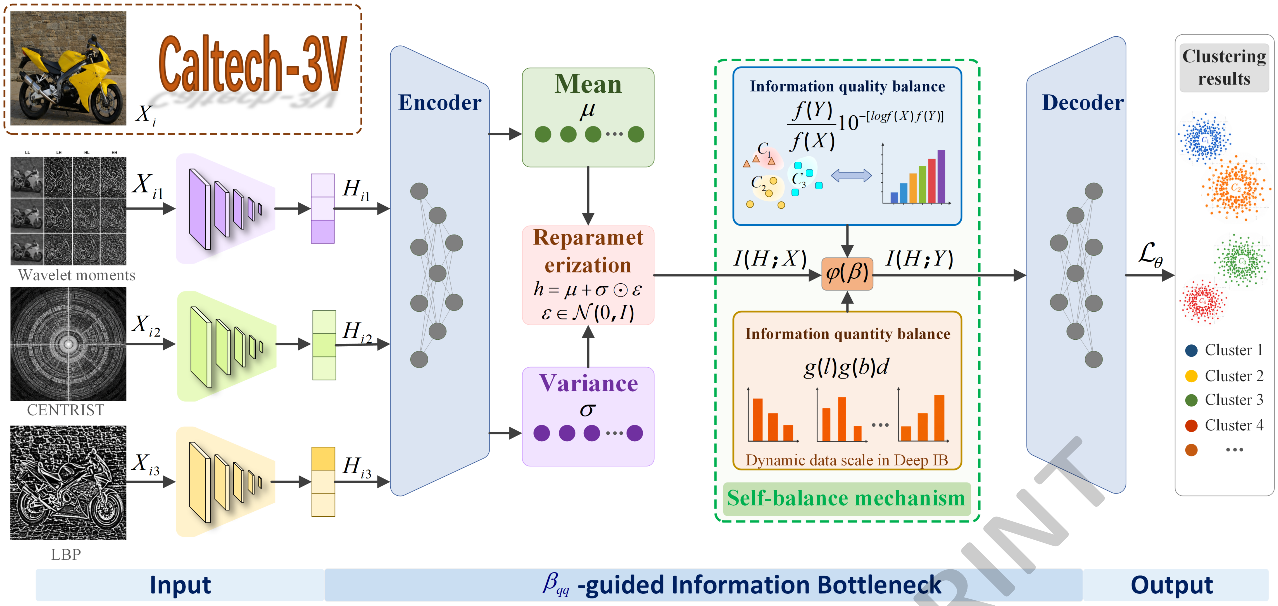

Figure 2: The overall architecture of our β_{qq} guided information bottleneck (β_{qq}IB) method. It empowers the self-balance of deep IB through by jointly modeling information-quality and information-quantity balance, enabling the learning of compact and meaningful representations for MMC.

tively learn both cluster-wise and modality-wise weights [Hu *et al.*, 2022]. (2) Deep IB method: Leveraging deep learning architectures to learn multimodal representation. For example, a contrastive learning framework is employed to explore cross-modality correlations by extracting cluster structures and modality features [Mao *et al.*, 2024]. Similarly, a transformer mechanism is employed to fuse features from different modalities in a way that benefits downstream tasks [Li *et al.*, 2025]. In previous work [Hu *et al.*, 2025b], deep variational technique is used to learn multi-aspect self-guided information from levles of modal features and cluster assignments.

However, all of the above methods encounter an intractable challenge of manually setting the trade-off operator β, which is central to IB theory for balancing information compression and preservation, as shown in Fig. 1. This manual strategy directly degrades the performance of IB-based MMC methods when dealing with surging volumes of multimodal data. In addition, manually setting β overlooks *information quality balance* and *information quantity balance*, leading to unsatisfactory balance of compressing task-irrelevant features and preserving task-relevant ones.

In this work, we propose a novel β_{qq}-guided information bottleneck (β_{qq} IB) for addressing the aforementioned problems, as shown in Fig. 2. The core of β_{qq} IB is a self-balance mechanism that consists of a $\beta_{quality}$ component and a $\beta_{quantity}$ component, which guides the proposed method to extract compact and meaningful features from different modalities. The $\beta_{quality}$ component aims to achieve an information quality balance by modeling the relation between task characteristics and sample scale, thereby facilitating the extraction of the most task-relevant features. Meanwhile, the $\beta_{quantity}$ component intends to realize information quantity balance through modeling the dynamic changes of data scale within the deep variational architect, enhancing the proposed method's adaptability to large-scale multimodal datasets. Experiments on six benchmark datasets demonstrate the superiority of β_{qq} IB method.

- To the best of our knowledge, we are the first to investigate self-balance learning in IB-based clustering.
- A novel β_{qq} IB method is proposed to enhance MMC performance in a β-free manner.
- A new self-balance mechanism is developed by integrating both information quality balance and information quantity balance.
- Experiments on diverse multimodal and large-scale datasets demonstrate the superiority of the β_{qq} IB method.

2 Related Work

2.1 Information Bottleneck

IB theory is a powerful mechanism for learning compact and meaningful representations by balancing mutual information minimization and maximization [Tishby *et al.*, 1999]. Recently, deep IB is proposed to encode data samples by a Gaussian latent distribution and optimize variational lower bounds through the reparameterization trick to achieve end-to-end training mode [Alemi *et al.*, 2017]. It has been successfully applied to a wide range of areas in computer vision and large language modeling, such as representation learning [Amjad and Geiger, 2020], video action recognition [Fan *et al.*, 2022], and information extraction [Wang *et al.*, 2025]. For MMC, IB theory reshapes representation learning into a non-redundant multimodal feature puzzle, thereby achieving promising clustering performance [Hu *et al.*, 2022; Yan *et al.*, 2024; Lou *et al.*, 2024; Lou *et al.*, 2025a; Mao *et al.*, 2025]. However, these MMC models treat the IB trade-off operator β as a hyperparameter and rely on manual tuning. This approach does not scale to the rapidly growing volume of multimodal data and overlooks the functional roles of information quality and information quantity in balancing information compression-preservation. In this work, we intend to address this challenge by empowering self-balance of deep IB to achieve satisfactory clustering results.

2.2 Multimodal Clustering

The MMC task aims to discover satisfactory cluster structures by fully exploiting complementary features across heterogeneous modalities. We classify existing MMC methods into two categories. (1) Non-IB-based methods: Learning rich multimodal representations with deep neural network architectures is important for identifying clustering [Zhang *et al.*, 2025; Zhao *et al.*, 2025]. For example, a graph neural network is constructed to obtain multimodal representation with rich high-level data features and non-Euclidean structural features [Li *et al.*, 2022]. Subsequently, Transformer mechanism is introduced to extract modality-specific discriminative features, aiming to address the feature conflict issue [Dong *et al.*, 2025]. The contrastive learning framework is proposed to learn enhanced multimodal representations to perform MMC task [Zou *et al.*, 2024; Lou *et al.*, 2025a]. (2) IB-based methods: Learning compact representations with IB theory has been an attractive approach. For instance, a differentiable IB framework is proposed to achieve representation compression without requiring variational approximation [Yan *et al.*, 2024]. More recently, contrastive learning has been combined with IB theory to learn compact and meaningful multimodal representations, leading to improved clustering performance [Lou *et al.*, 2025b]. The sufficiency criterion is employed to reduce cross-modality redundancy and learn compact multimodal representation [Mao *et al.*, 2025].

3 The Proposed β_{qq} IB Method

3.1 Problem Formulation

Let $X = \{X_1, X_2, ..., X_n\}$ be the input data samples, where $X_i = \{x_{ij}\}_{j=1}^{j=m}$ represents the different modalities for each sample. $H = \{H_1, H_2, ..., H_n\}$ denotes multimodal representation, which fuses features from each modality. $\{Y_i\}_{i=1}^{i=n}$ represents the clustering assignment of multimodal representation H classified by the clustering model. The objective of β_{qq} IB is to learn a compact and meaningful representation $p(h|x)$ of X to Y in self-balance paradigm, where C is the number of clusters.

3.2 Overview of Objective Function

In this part, we propose a novel self-balance mechanism to address the issue of manually setting β in IB-based MMC methods, with the overall objective function defined as follows:

$$\mathcal{L}_\theta = I(H;X) - \varphi(\beta)I(H;Y) \tag{1}$$

where φ is a self-balance mechanism that eliminates the need for manual configuration and θ denotes a set of hyperparameters.

3.3 Self-balance Mechanism

In this section, a novel self-balance mechanism is proposed for β_{qq} IB method, where the self-balance learning consists of two parts, including information quality balance and information quantity balance. At its core, it reflects dynamic–static dependency, a principle that resonates with the function of the dam crest in hydraulic engineering. The overall self-balance learning is defined as:

$$\varphi(\beta) = \beta_{quality} + \beta_{quantity} \tag{2}$$

where $\beta_{quality}$ denotes the information quality balance and $\beta_{quantity}$ is information quantity balance.

Information Quality Balance. This component models the relation between task characteristics and sample scale to balance mutual information compression and preservation, aiming to extract the most task-relevant features. The detailed formulation is provided as follows:

$$\beta_{quality} = \frac{f(Y)}{f(X)} 10^{-[log f(X) f(Y)]} \tag{3}$$

where $f(\cdot)$ is the spatial size of the discrete object.

The above formula balances the relative importance of data samples X and labels Y to ensure that the cluster information is not overshadowed by the data sample information. The first term of Eq. (3) is used to quantify the relative importance of cluster information to maintain the proposed model can prioritize more important task-relevant information. The second term of Eq. (3) adjusts information compression by applying a logarithmic transformation to the joint information between data samples and labels, which intends to preserve better quality of information.

Information Quantity Balance. In this part, we model the dynamic changes of data scale within the deep variational architecture to achieve information quantity balance, aiming to enhance adaptability of the proposed method. It works by using the hyperparameters in θ to balance information quantity, with the detailed formulation provided as follows:

$$\beta_{quantity} = g(l)g(b)d \tag{4}$$

where $g(\cdot)$ denotes the root operation, $\{l, b, d\} \in \theta$ corresponds to the learning rate, batchsize, and dropout rate, respectively.

The above formula represents the automatic adjustment of information distribution through the interaction of learning rate, batchsize, and dropout rate, thereby balancing information compression-preservation of the proposed method. This mechanism is consistent with the statistical averaging principle underlying the law of large numbers: under large-scale optimization dynamics, the overall distribution of task-relevant and task-irrelevant information tends to exhibit approximate statistical stability. Therefore, this strategy ensures that the model can effectively learn from large-scale data while preventing overfitting and information loss, optimizing the balance of information throughout the learning process.

Algorithm 1 Algorithm for Optimizing the β_{qq} IB

1: **Input:** Multimodal dataset X, cluster number C, hyperparameter set $\{l, b, d\}$
2: **Output:** Final clustering results
3: **Initialize:** $p(x, y) \leftarrow$ Calculate joint probability distributions
4: $\varphi(\beta) \leftarrow$ Learn self-balance trade-off operator via Eq. (2)
5: **repeat**
6: $H \leftarrow$ adopt deep variational techniques to encode multimodal representation
7: Compute the $\mathrm{CrossEntropy}(y, \hat{y})$ between pseudo-label and learned label
8: $\mathbb{D}(p(H|X), q(H)) \leftarrow$ according to the Gaussian distribution $\mathcal{N}(\mu(x), \sigma(x))$
9: $\mathcal{L} \leftarrow$ compute loss of Eq. (1)
10: Utilize AdamW algorithm to optimize the $\mathcal{L}$
11: **until** Convergence

3.4 Analysis on the Method

Observing from the proposed β_{qq} IB method, there are three main advantages compared with IB-based MMC methods.

- To the best of our knowledge, this is the first work to investigate self-balance in IB-based MMC methods, solving the long-standing intractable issue of manually setting the trade-off operator β.
- The novel self-balance mechanism incorporates information quality balance and information quantity balance, leading to more compact representations and improved clustering results.
- The β_{qq} IB operates as a β-free paradigm, making it well-suited for use in a variety of real-world applications.

3.5 Optimization

Following previous work [Alemi *et al.*, 2017; Wang *et al.*, 2025], we adopt gradient descent algorithm to optimize Eq. (1) in an end-to-end fashion. It aims to achieve promising clustering results through a self-balance pattern. The detailed algorithmic framework is presented in Algorithm 1.

4 Experimental Validation

Dataset	Modality	Sample	Clusters
Caltech_3V	3	1400	7
Caltech_4V	4	1400	7
CCV	3	6773	20
Event	3	1579	8
Large Dataset	Modality	Sample	Clusters
ESP-Game	4	11032	7
MIRFlickr	4	12154	6

Table 1: Details of the involved datasets.

Method	Value Space
DIB	$\{10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$
MNRC	$\{0.3, 0.6, 0.9, 1.1.2, 1.5, 1.8\}$
SDCIB	$\{10^0, 10, 10^2, 10^3, 10^4\}$
MSDIB	$\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$
β_{qq} IB	β-free

Table 2: β settings of IB-based methods.

4.1 Datasets

We used six publicly available multimodal datasets as well as large-scale datasets to validate the superiority of the proposed β_{qq} IB method. The details are presented in Tab. 1.

Caltech[1] is an image dataset consisting of 1,400 image samples categorized into 7 clusters, which derives two multimodal datasets **Caltech_3V** and **Caltech_4V**. The Caltech_3V is constructed from three modalities: Wavelet moments, CENTRIST, and LBP. The Caltech_4V adds a new GIST modality on top of Caltech_3V.

CCV[2] is a video dataset containing 6,673 samples from 20 clusters, featuring three types of views: SIFT, STIP, and MFCC.

Event[3] contains eight kinds of sports event classes with 1579 images, which includes three modalities of shape, color and texture.

ESP-Game[4] is a social image dataset collected from an online image annotation game, consisting of 11,032 social image instances, divided into 7 clusters.

MIRFlickr[5] is a cross-modal dataset consisting of images and corresponding labels, containing 12,154 instances divided into 6 clusters.

4.2 Compared Methods

We compared the proposed method with various state-of-the-art MMC methods, categorized into two types:

Non-IB-based MMC methods: These methods mainly refer to the use of classic machine learning algorithms and deep learning frameworks to identify clustering patterns, in which **K-means** [McQueen, 1967], **NCuts** [Shi and Malik, 2000], **SMVSC** [Sun *et al.*, 2021], **SPDMVC** [Chen *et al.*, 2024], **DIVIDE** [Lu *et al.*, 2024], **LUCE-CMC** [Zou *et al.*, 2024], **ICMVC** [Chao *et al.*, 2024], **PGCL-DCL** [Hu *et al.*, 2025c], and **STCMVC** [Hu *et al.*, 2025b], **PDMC-RCL** [Lou *et al.*, 2025a].

IB-based MMC methods: We select four state-of-the-art IB-based MMC methods for experimental comparison, in which **DIB** [Yan *et al.*, 2024], **MNRC** [Mao *et al.*, 2025], **SDCIB** [Lou *et al.*, 2025b], and **MSDIB** [Hu *et al.*, 2025a]. Tab. 2 demonstrates the β values used by the aforementioned

[1]https://data.caltech.edu/records/mzrjq-6wc02
[2]https://www.ee.columbia.edu/ln/dvmm/CCV/
[3]http://vision.stanford.edu/lijiali/event_dataset/
[4]https://www.kaggle.com/datasets/parhamsalar/espgame?utm_source=chatgpt.com
[5]http://press.liacs.nl/mirflickr/

Method	Caltech_3V		Caltech_4V		CCV		Event		ESP-Game		MIRFlickr	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Kmeans	46.3	31.3	54.6	46.7	19.4	17.6	34.7	20.7	35.4	21.8	38.1	25.6
Ncuts (TPAMI'00)	42.6	25.4	67.8	47.9	21.6	17.8	34.8	15.5	37.2	22.9	46.1	29.8
SMVSC (MM'21)	54.8	43.3	60.1	53.3	20.6	16.3	49.0	32.1	56.0	39.0	53.4	33.3
SPDMC (TNNLS'23)	51.4	40.5	66.5	60.9	10.8	6.0	30.2	13.6	39.1	21.4	58.4	36.6
DIVIDE (AAAI'24)	60.9	53.8	64.3	57.9	16.4	11.1	33.9	16.2	41.1	20.0	52.6	30.7
LUCE-CMC (MM'24)	75.6	62.9	80.6	70.0	34.3	32.4	58.7	42.8	51.6	34.7	**60.4**	**42.0**
ICMVC (AAAI'24)	58.2	43.0	83.1	72.2	10.9	6.6	36.4	30.3	45.8	29.5	43.5	24.4
PGCL-DCL (TNNLS'24)	71.9	61.5	71.6	62.7	35.5	32.7	51.6	33.5	47.4	30.0	58.0	36.8
STCMVC (AAAI'25)	**76.3**	**71.6**	86.9	78.8	22.1	25.4	56.9	40.1	56.0	38.0	57.0	40.6
PDMC-RCL (TIP'25)	67.9	57.6	85.1	75.9	14.1	8.8	19.4	2.4	35.0	15.3	50.4	31.4
DIB (CVPR'24)	73.0	63.9	75.9	74.9	23.1	24.8	41.7	31.2	59.1	37.8	55.3	39.1
MNRC (TII'25)	62.4	58.1	73.4	66.5	16.9	15.3	50.2	37.0	58.7	44.5	29.2	14.9
SDCIB (ICML'25)	70.8	62.8	81.8	71.0	34.3	31.7	60.1	43.7	52.3	36.7	54.7	39.2
MSDIB (AAAI'25)	66.1	56.5	71.1	63.9	22.7	18.9	30.6	12.4	47.4	37.7	58.4	37.5
β_{qq} IB	72.7	63.3	**88.7**	**80.7**	**36.3**	**35.5**	**62.2**	**45.6**	**68.1**	**45.5**	<u>59.3</u>	<u>40.9</u>

Table 3: Clustering performance comparison on multimodal datasets. Bold values indicate the best results. Underline values represent the second-best results.

Method	Caltech_3V		Caltech_4V		CCV		Event		ESP-Game		MIRFlickr	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
DIB	73.0	63.9	75.9	74.9	23.1	24.8	41.7	31.2	35.8	21.8	55.3	39.1
DIB + β_{qq}	74.6	65.4	84.2	76.2	27.0	27.6	49.8	35.1	40.5	26.8	56.6	40.9
MNRC	62.4	58.1	73.4	66.5	16.9	15.3	50.2	37.0	58.7	44.5	29.2	14.9
MNRC + β_{qq}	63.5	58.6	73.2	68.8	18.3	18.3	51.1	37.3	60.6	47.1	34.8	16.4
SDCIB	70.8	62.8	81.8	71.0	34.3	31.7	60.1	43.7	52.3	36.7	59.1	41.5
SDCIB + β_{qq}	72.7	63.3	88.7	80.7	36.3	35.5	62.2	45.6	68.1	45.5	59.3	40.9
MSDIB	66.1	56.5	71.1	63.9	22.7	18.9	30.6	12.4	47.4	32.1	58.4	37.5
MSDIB + β_{qq}	70.0	62.3	77.2	70.8	20.2	19.1	31.0	14.2	52.5	37.7	58.7	40.8
β_{qq} IB	72.7	63.3	88.7	80.7	36.3	35.5	62.2	45.6	68.1	45.5	59.3	40.9

Table 4: Ablation study on the multimodal datasets. All IB-based methods exhibit consistent gains in ACC and NMI when equipped with the proposed β_{qq} self-balance mechanism.

IB-based MMC methods. Our proposed β_{qq} IB is β-free, as the trade-off operator is automatically learned by the proposed self-balance mechanism.

4.3 Settings of Experiments

All experiments were conducted using the PyTorch framework on a server equipped with an Intel(R) Xeon(R) Gold 6430 CPU and four NVIDIA RTX 4090 GPUs, each with 24 GB of memory, running Ubuntu 22.04. We follow the data encoding strategy adopted in prior work [Lou *et al.*, 2025b]. A grid search strategy was employed to identify the optimal learning rate from $\{10^{-2}, 10^{-3}, 10^{-4}\}$, the optimal batch size from $\{100, 200\}$, and the optimal dropout rate from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. The β value for each set of experiments was automatically calculated via Eq. (2).

For a more thorough analysis, we use two popular metrics – Cluster Accuracy (ACC) and Normalized Mutual Information (NMI) to evaluate the clustering performance. And a higher value indicates better model performance. For all methods, we perform 20 runs on each dataset with 100 epochs per run, and select the best metrics from the 20 runs.

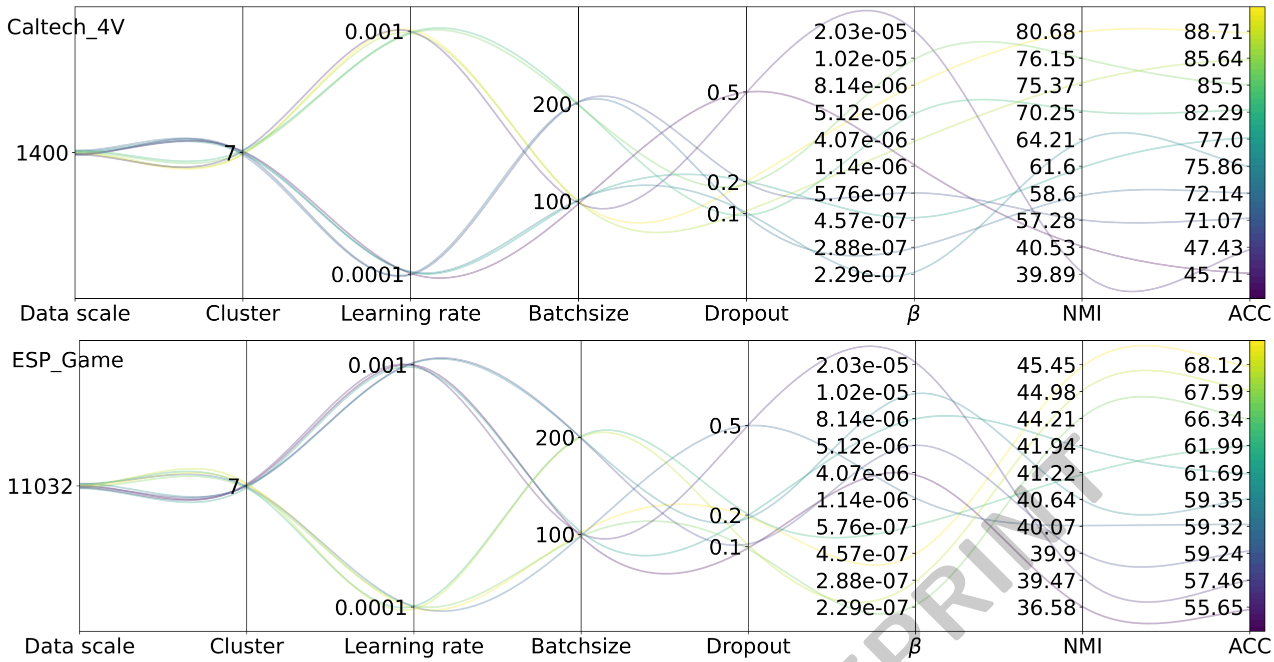

Figure 3: Parameter sensitivity analysis of the proposed β_{qq} IB on Caltech_4V and ESP-Game datasets.

4.4 Analysis on MMC Results

In this part, we conduct a series of comprehensive experiments to compare our β_{qq} IB method with various state-of-the-art MMC methods. Tab. 3 reports the comparison of clustering performance on six benchmark datasets.

The β_{qq} IB method achieves competitive clustering results compared to Non-IB-based methods, highlighting the advantages of information compression-preservation mechanism in IB theory. It enables the proposed method to learn compact multimodal representation by compressing task-irrelevant features. For example, the proposed method outperforms the best SMVSC method on the ESP dataset, achieving scores of 12.1% and 6.5% on the ACC and NMI metrics, respectively. In addition, β_{qq} IB method consistently outperforms the clustering results of IB-based methods across all multimodal datasets. For instance, our method outperforms the best PDMC-RCL baseline by 4.8% in NMI on the Caltech_4V dataset. This clearly suggests that the self-balance mechanism guides the proposed method to achieve an optimal trade-off in identifying clustering patterns. The relatively modest performance of the β_{qq} IB method on the Caltech_3V dataset may be attributed to the fact that its three modalities mainly emphasize local features, resulting in limited modal complementarity. Furthermore, the satisfactory clustering results on the large-scale ESP-Game and MIR-Flickr datasets demonstrate the robustness of the proposed method. Building on the rate distortion principle of IB theory, the β_{qq}-IB algorithm adaptively balances information compression and preservation in a β-free manner. This improves the practicality of the β_{qq}-IB algorithm in real-world application scenarios.

4.5 Ablation Study

To verify the effectiveness of the self-balance mechanism, we conducted ablation experiments on all IB-based methods. The experimental results in Tab. 4 show that incorporating the automatically learned trade-off operator β (referred to as β_{qq}) consistently improves performance of IB-based methods. For instance, with our self-balance mechanism, the DIB method improves ACC on the Event dataset by 8.1%. This is because the proposed self-balance mechanism automatically learns β by jointly accounting for information quality balance and information quantity balance, leading to compact multimodal representation and improved clustering results. It is worth noting that our model incorporates the self-balance mechanism into the SDCIB framework, making the results of β_{qq} IB consistent with those of SDCIB+β_{qq}.

4.6 Parameter Sensitivity Analysis

To further analyze the stability and robustness of the proposed β_{qq} IB method, we conduct a parameter sensitivity analysis on the Caltech_4V and ESP-Game datasets, as shown in Fig. 3. Given the variance in datasets and hyperparameters, the self-balance mechanism adjusts the β values to guide the proposed method toward optimal representations. Notably, self-balance mechanism is fundamentally different from parameter transfer. It eliminates the for manual tuning of the β parameter, which provides valuable insights into the generalization of IB methods. Furthermore, it is an adaptive approach that collaborate with hyperparameters to mine clustering patterns from the perspectives of information-quality and information-quantity balance.

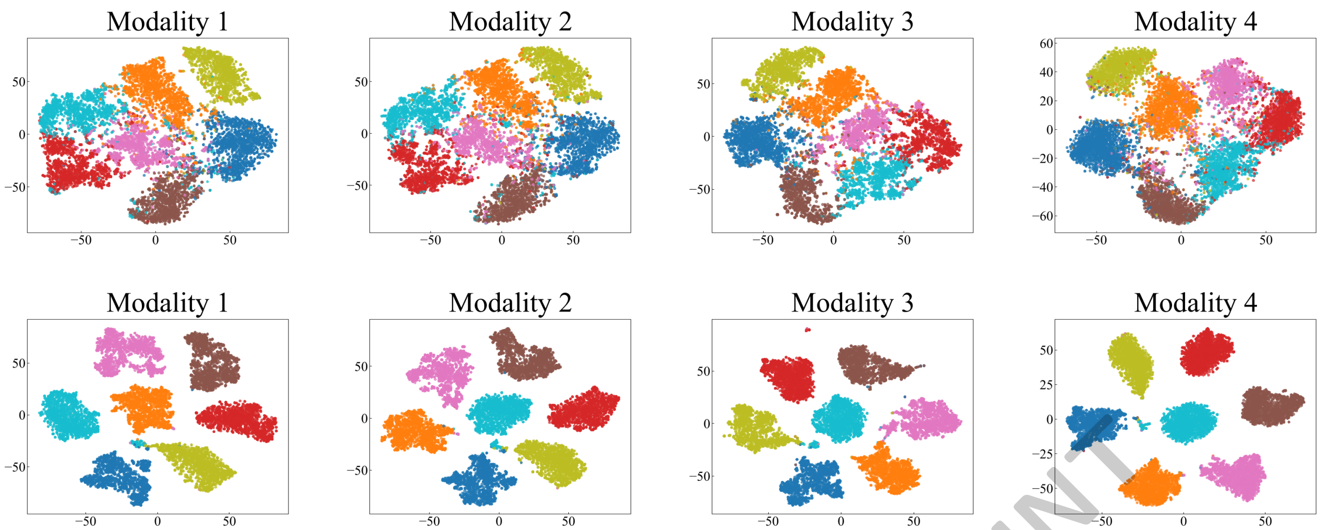

Figure 4: The t-SNE visualization comparing the clustering performance on the ESP-Game dataset. The top and bottom images respectively show results of SDCIB and β_{qq} IB.

4.7 Visualization Analysis

We select the ESP-Game dataset for clustering visualization using the popular t-SNE method to further illustrate the clustering results of the β_{qq} IB method. Fig. 4 shows the clustering results of SDCIB (the latest IB-based MMC) and the proposed method. Our β_{qq} IB method results in a more compact cluster structure and clearer separation across different clusters compared to the SDCIB method. This observation further reflects that our β_{qq} IB method achieves satisfactory clustering results, due to the novel and effective self-balance mechanism that empowers the self-balance of deep IB from a more objective and fine-grained perspective.

5 Conclusion

In this work, we propose a novel β_{qq} IB method to address the long-standing challenge of manually setting the trade-off parameter β in IB-based MMC methods. We introduce a self-balance mechanism by taking into accounting with both information quality balance and information quantity balance. The information quality balance models the relation between task characteristics and data sample scale, ensuring that cluster-relevant features are not overshadowed by data sample information. Meanwhile, the information quantity balance models the dynamic changes in sample scale within the deep variational architecture, enhancing the robustness of the proposed method. Experiments on six multimodal datasets as well as large-scale datasets demonstrate the effectiveness of the proposed method. However, several challenges remain when extending the proposed method to real-world multimodal applications. For example, incomplete modality information may weaken the fused representation, thereby adversely affecting clustering performance. Moreover, the deep variational architecture underlying the proposed method imposes certain computational demands, which are necessary for stable optimization and effective representation learning.

In the future, we focus on the propagation-balance version of the IB-based MMC methods [Hu *et al.*, 2024], and extend to multimodal information extraction [Liu *et al.*, 2026] and multimodal large language model [Sapkota and Karkee, 2026].

Ethical Statement

This work does not raise ethical concerns, as it relies solely on public datasets and does not involve human subjects or sensitive data.

Acknowledgements

This work is supported by the National Natural Science Foundation of China [under Grant No.62576320], in part by Henan Province Outstanding Youth Science Fund Program [under Grant 252300421223], in part by Natural Science Foundation of Henan Province [under Grant No.252300423339].

References

[Alemi *et al.*, 2017] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[Amjad and Geiger, 2020] Rana Ali Amjad and Bernhard C. Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(9):2225–2239, 2020.

[Chao *et al.*, 2024] Guoqing Chao, Yi Jiang, and Dianhui Chu. Incomplete contrastive multi-view clustering with high-confidence guiding. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 11221–11229, 2024.

[Chen *et al.*, 2024] Rui Chen, Yongqiang Tang, Yuan Xie, Wenlong Feng, and Wensheng Zhang. Semisupervised progressive representation learning for deep multiview clustering. *IEEE Trans. Neural Networks Learn. Syst.*, 35(10):14341–14355, 2024.

[Chowdhury *et al.*, 2025] Anal Roy Chowdhury, Avisek Gupta, and Swagatam Das. Deep multi-view clustering: A comprehensive survey of the contemporary techniques. *Inf. Fusion*, 119:103012, 2025.

[Dong *et al.*, 2025] Zhibin Dong, Meng Liu, Siwei Wang, Ke Liang, Yi Zhang, Suyuan Liu, Jiaqi Jin, Xinwang Liu, and En Zhu. Enhanced then progressive fusion with view graph for multi-view clustering. In *Proceedings of the 42nd IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15518–15527, 2025.

[Fan *et al.*, 2022] Jiawei Fan, Yu Zhao, Xie Yu, Lihua Ma, Junqi Liu, Fangqiu Yi, and Boxun Li. DTR: an information bottleneck based regularization framework for video action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3877–3885, 2022.

[Fang *et al.*, 2023] Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Trans. Knowl. Data Eng.*, 35(12):12350–12368, 2023.

[Hu *et al.*, 2022] Shizhe Hu, Zhengzheng Lou, and Yangdong Ye. View-wise versus cluster-wise weight: Which is better for multi-view clustering? *IEEE Trans. Image Process.*, 31:58–71, 2022.

[Hu *et al.*, 2024] Shizhe Hu, Zenglin Shi, Xiaoqiang Yan, Zhengzheng Lou, and Yangdong Ye. Multiview clustering with propagating information bottleneck. *IEEE Trans. Neural Networks Learn. Syst.*, 35(7):9915–9929, 2024.

[Hu *et al.*, 2025a] Shizhe Hu, Jiahao Fan, Guoliang Zou, and Yangdong Ye. Multi-aspect self-guided deep information bottleneck for multi-modal clustering. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pages 17314–17322, 2025.

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

[Hu *et al.*, 2025c] Shizhe Hu, Chengkun Zhang, Guoliang Zou, Zhengzheng Lou, and Yangdong Ye. Deep multiview clustering by pseudo-label guided contrastive learning and dual correlation learning. *IEEE Trans. Neural Networks Learn. Syst.*, 36(2):3646–3658, 2025.

[Li *et al.*, 2022] Xuelong Li, Hongyuan Zhang, and Rui Zhang. Adaptive graph auto-encoder for general data clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):9725–9732, 2022.

[Li *et al.*, 2025] Yuzhe Li, Haocheng Sun, Jiayi Cai, and Jin Wu. Mvib-lip: Multi-view information bottleneck for visual speech recognition via time series modeling. *Entropy*, 27(11):1121, 2025.

[Liu *et al.*, 2026] Jiang Liu, Bobo Li, Xinran Yang, Na Yang, Hao Fei, Mingyao Zhang, Fei Li, and Donghong Ji. M3d: A multimodal, multilingual and multitask dataset for grounded document-level information extraction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(1):807–823, 2026.

[Lou *et al.*, 2024] Zhengzheng Lou, Chaoyang Zhang, Hang Xue, Yangdong Ye, Qinglei Zhou, and Shizhe Hu. Self-supervised weighted information bottleneck for multi-view clustering. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 4643–4650. ijcai.org, 2024.

[Lou *et al.*, 2025a] Zhengzheng Lou, Hang Xue, Yanzheng Wang, Chaoyang Zhang, Xin Yang, and Shizhe Hu. Parameter-free deep multi-modal clustering with reliable contrastive learning. *IEEE Trans. Image Process.*, 34:2628–2640, 2025.

[Lou *et al.*, 2025b] Zhengzheng Lou, Ke Zhang, Yucong Wu, and Shizhe Hu. Super deep contrastive information bottleneck for multi-modal clustering. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[Lu *et al.*, 2024] 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 38th AAAI Conference on Artificial Intelligence*, pages 14193–14201, 2024.

[Mao *et al.*, 2024] Yiqiao Mao, Xiaoqiang Yan, Shizhe Hu, and Yangdong Ye. Contrastive cross-modal clustering with twin network. *Pattern Recognit.*, 155:110645, 2024.

[Mao *et al.*, 2025] Yiqiao Mao, Xiaoqiang Yan, Qiang Guo, and Yangdong Ye. Multimodal nonredundant clustering via sufficiency complementary mining. *IEEE Trans. Ind. Informatics*, 21(2):1872–1881, 2025.

[McQueen, 1967] James McQueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[Sapkota and Karkee, 2026] Ranjan Sapkota and Manoj Karkee. Object detection with multimodal large vision-language models: An in-depth review. *Inf. Fusion*, 126:103575, 2026.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[Sun *et al.*, 2021] Mengjing Sun, Pei Zhang, Siwei Wang, Sihang Zhou, Wenxuan Tu, Xinwang Liu, En Zhu, and Changjian Wang. Scalable multi-view subspace clustering with unified anchors. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3528–3536, 2021.

[Tishby and Slonim, 2000] Naftali Tishby and Noam Slonim. Data clustering by markovian relaxation and the information bottleneck method. In *Proceedings of the 14th Advances in Neural Information Processing Systems*, pages 640–646, 2000.

[Tishby *et al.*, 1999] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication Control Computing*, pages 368–377, 1999.

[Wang *et al.*, 2025] Youwei Wang, Peisong Cao, Haichuan Fang, and Yangdong Ye. Span-aware pre-trained network with deep information bottleneck for scientific entity relation extraction. *Neural Networks*, 186:107250, 2025.

[Yan *et al.*, 2020] Xiaoqiang Yan, Yiqiao Mao, Shizhe Hu, and Yangdong Ye. Heterogeneous dual-task clustering with visual-textual information. In Carlotta Demeniconi and Nitesh V. Chawla, editors, *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 658–666, 2020.

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

[Zhang *et al.*, 2025] Peiliang Zhang, Jingling Yuan, Chao Che, Yongjun Zhu, and Lin Li. Subgraph information bottleneck with causal dependency for stable molecular relational learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence,*, pages 7931–7939. ijcai.org, 2025.

[Zhao *et al.*, 2025] Liang Zhao, Ziyue Wang, Chuanye He, Qingchen Zhang, and Bo Xu. Dual robust unbiased multi-view clustering for incomplete and unpaired information. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pages 7092–7100. ijcai.org, 2025.

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