WHAT TO PRESERVE AND TRUST IN MULTI-MODAL CLUSTERING: CROSS-GUIDED INFORMATION BOTTLENECK WITH RELIABLE CONTRASTIVE LEARNING

Paper ID: 19731

ABSTRACT

Multi-modal clustering (MMC) aims to discover common cluster structures from multiple modalities. Effective MMC requires determining what information to preserve and how strongly negative relations are trusted. Existing information bottleneck (IB)-based MMC methods can suppress unnecessary information while preserving shared content, but fail to explicitly consider that the contribution of each modality may vary across samples. Moreover, existing contrastive MMC methods can promote cross-modal consistency, but often overlook that negative pairs may contribute differently, especially when their relations are ambiguous or the samples involved are uncertain or inconsistent across modalities. To address these issues, we propose a novel reliable multi-modal clustering framework named CGIB-RCL, consisting of Cross-Guided Information Bottleneck (CGIB) and Reliable Contrastive Learning (RCL). Specifically, CGIB constructs guidance for each target modality by adaptively combining the remaining modalities for each sample, preserving information supported across modalities while reducing modality-specific content. RCL separately models relation ambiguity and sample reliability to determine the contribution of each negative pair, reducing the influence of ambiguous relations and pairs involving unreliable samples during contrastive learning. Together, CGIB and RCL regulate what information is preserved and how strongly negative relations are trusted during learning. Extensive experiments on five datasets demonstrate the effectiveness of the proposed framework. The source code will be available on the author's GitHub homepage upon acceptance.

1 INTRODUCTION

Multi-modal clustering (MMC) aims to discover a common cluster structure from data observed through multiple modalities. Different modalities capture the same sample from different perspectives and often provide complementary information. However, not all information in a modality is equally useful for discovering the common structure. Some information may be specific to one modality and less relevant to the clustering task. At the same time, cross-modal learning often relies on positive and negative relations between samples, while negative relations may be unreliable in unsupervised settings. This raises two related questions: ***What information should be preserved in the learned representations?*** and ***How strongly should negative relations be trusted during contrastive learning?***

Information Bottleneck (IB) provides a natural framework for controlling the information retained in a representation (Tishby et al., 2000; Alemi et al., 2017). Its basic principle is to preserve useful information while suppressing unnecessary content. MIB (Federici et al., 2020) extends this idea to unsupervised multi-modal learning by preserving information shared between modalities while reducing modality-specific content. IB-based MMC methods such as SUMVC (Cui et al., 2023) and STMVC (Ren et al., 2026) further exploit sufficient cross-modal information to learn consistent representations. More recent methods also consider differences among modalities. PTIB (Lou et al., 2025b) uses the remaining modalities to evaluate a target modality, OMIB (Wu et al., 2025) dynamically adjusts modality-wise regularization, and OVA-IB (Li et al., 2026) characterizes each modality relative to the others. These methods account for modality differences in different ways, but modality-level adaptation may not fully capture how the usefulness of a supporting modality

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

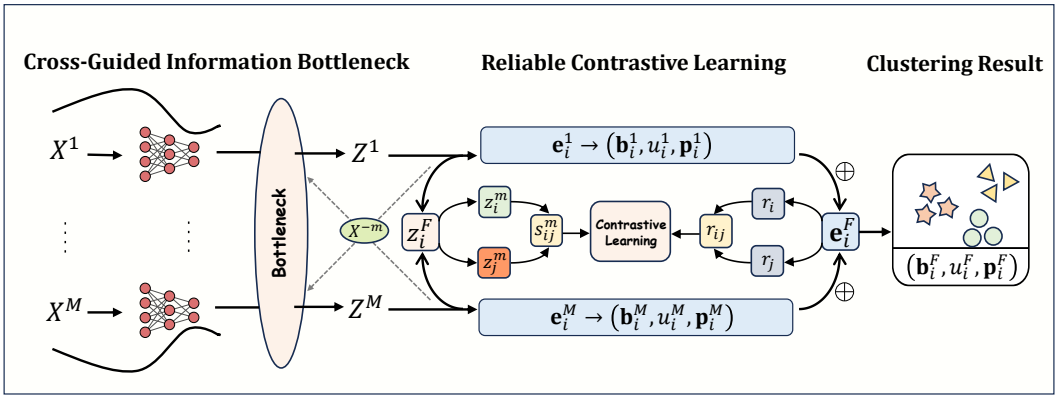

Figure 1: Overview of the proposed CGIB-RCL framework. First, each modality is encoded into a latent representation. Next, for each target modality, CGIB constructs sample-specific guidance from the supporting modalities to retain cross-modal information while reducing modality-specific content. Then, the representations are used to obtain relation scores, while evidential fusion is used to estimate pair reliability. RCL uses these two factors to determine how strongly each negative pair is trusted.

varies across samples. A modality that provides useful information for one sample may be less informative for another, motivating control over information preservation at the sample level.

Contrastive MMC methods provide another important mechanism for learning cross-modal consistency. CoMVC (Trosten et al., 2021), MFLVC (Xu et al., 2022), DeepMVC (Trosten et al., 2023), DealMVC (Yang et al., 2023), and CVCL (Chen et al., 2023) use contrastive objectives to improve alignment or consistency across modalities at different representation levels. In these methods, representations of the same sample are typically treated as positive pairs, while different samples form negative pairs. However, different samples do not necessarily belong to different clusters, so samples from the same underlying cluster may be incorrectly treated as negatives (Hu et al., 2025b; Lu et al., 2024). Recent studies have refined sample relations or improved reliability from different perspectives. Potential false negatives can be filtered or refined through evolving clustering information and high-order relations (Hu et al., 2025b; Lu et al., 2024; Wang & Feng, 2024), while other methods reduce the effect of noisy modalities, model predictive uncertainty, identify unreliable observations, or adapt cross-modal alignment (Xu et al., 2024a; Hu et al., 2025a; Yang et al., 2025; Lou et al., 2025a; Peng et al., 2026). Nevertheless, negative pairs may still contribute differently. The relation itself can be ambiguous when two samples are highly similar, while the samples forming the pair can also be uncertain or inconsistent across modalities. These two factors characterize different aspects of a negative pair and should be considered separately when determining its contribution to contrastive learning.

Recent MMC studies have also combined IB with contrastive learning. MSCIB (Yan et al., 2024) uses contrastive semantic consistency to enhance information bottleneck learning across views. IBFDMVC (Hu et al., 2024) employs contrastive learning to enhance representations and information bottleneck learning to reduce representation noise, while SDCIB (Lou et al., 2025c) combines information bottleneck learning with contrastive objectives over hidden representations and clustering assignments. These studies suggest that IB and contrastive learning are complementary in MMC, with IB controlling the information retained in representations and contrastive learning shaping representations through cross-modal sample relations. However, their combination alone does not determine what information should be preserved for each sample or how strongly different negative relations should be trusted during learning.

To address these issues, we propose CGIB-RCL, a novel reliable multi-modal clustering framework with Cross-Guided Information Bottleneck (CGIB) and Reliable Contrastive Learning (RCL), as illustrated in Figure 1. The framework jointly determines what information is preserved and how strongly negative relations are trusted through sample-specific cross-modal guidance and adaptive weighting of negative pairs. Specifically, CGIB addresses the first question by forming sample-specific guidance from the supporting modalities, encouraging the target representation to retain

information supported across modalities while reducing modality-specific content. RCL addresses the second question by weighting negative pairs according to relation ambiguity and sample reliability, reducing the influence of ambiguous relations and pairs involving unreliable samples during contrastive learning. Together, these two components allow CGIB-RCL to preserve information supported across modalities while reducing the influence of unreliable negative relations, addressing the two questions within a unified framework.

Our main contributions are summarized as follows:

- We propose CGIB-RCL, a new reliable multi-modal clustering framework that jointly addresses what information to preserve and how strongly negative relations are trusted during learning.
- We develop Cross-Guided Information Bottleneck (CGIB) to address what information to preserve. CGIB constructs sample-specific cross-modal guidance by adaptively weighting the supporting modalities, thereby preserving information supported across modalities while suppressing modality-specific content.
- We introduce Reliable Contrastive Learning (RCL) to address how strongly negative relations should be trusted. RCL separately models relation ambiguity and sample reliability to adaptively weight negative pairs, reducing the influence of ambiguous relations and pairs involving unreliable samples during contrastive learning.

2 Related Work

2.1 Information Bottleneck in Multi-modal Clustering

Information Bottleneck provides a principle for learning compact representations by preserving relevant information while suppressing unnecessary content (Tishby et al., 2000; Alemi et al., 2017). Existing multi-modal studies have explored shared or sufficient cross-modal information (Federici et al., 2020; Cui et al., 2023; Ren et al., 2026), information bottleneck learning combined with contrastive objectives (Yan et al., 2024; Hu et al., 2024; Lou et al., 2025c), and information regulation across modalities (Lou et al., 2025b; Wu et al., 2025; Li et al., 2026). Encouraging cross-modal dependence alone does not necessarily remove modality-specific information (Almudévar et al., 2025). Building on these studies, CGIB focuses on how the contributions of multiple supporting modalities vary across samples and uses sample-specific weights to construct guidance for each target modality.

2.2 Contrastive and Reliable Multi-modal Clustering

Contrastive learning is widely used in MMC to encourage consistency across modalities (Trosten et al., 2021; Xu et al., 2022; Trosten et al., 2023; Yang et al., 2023; Chen et al., 2023). Recent methods further improve the quality of contrastive relations by filtering or refining potential false negatives (Hu et al., 2025b; Lu et al., 2024; Wang & Feng, 2024), while other methods address reliability by reducing the effect of noisy modalities, modeling sample uncertainty, identifying unreliable observations, or adapting cross-modal alignment (Xu et al., 2024a; Hu et al., 2025a; Yang et al., 2025; Lou et al., 2025a; Peng et al., 2026). RCL differs by explicitly separating relation ambiguity from sample reliability when determining the influence of each negative pair. Relation ambiguity reflects how clearly a negative relation can be distinguished from its corresponding matched relation (Xu et al., 2024b), while sample reliability is estimated separately from predictive uncertainty and cross-modal consistency.

3 Method

3.1 Problem Definition

Given a multi-modal dataset $\mathcal{X} = \{x_1, \ldots, x_N\}$ with N samples, each sample $x_i = \{x_i^1, \ldots, x_i^M\}$ consists of M modalities, where $x_i^m \in \mathbb{R}^{d_m}$ denotes the m-th modality of sample i. Multi-modal clustering aims to partition the N samples into K clusters. For sample i, a modality-specific encoder parameterizes a latent distribution $q_m(z \mid x_i^m)$, from which a latent representation $z_i^m \in \mathbb{R}^d$ is

sampled. CGIB-RCL aims to learn discriminative and reliable representations by determining what information to preserve and how strongly negative relations should be trusted during contrastive learning.

3.2 Framework Overview

Figure 1 illustrates the overall framework of CGIB-RCL, which consists of Cross-Guided Information Bottleneck (CGIB) and Reliable Contrastive Learning (RCL). Each modality is first encoded into a latent representation z_i^m. For each target modality m, CGIB uses the remaining modalities as supporting modalities X^{-m} and estimates their sample-specific contributions to construct cross-modal guidance. The guidance regulates the target representation to preserve information supported across modalities while suppressing modality-specific content.

To assess whether the representations of different samples z_i^m and z_j^m can be clearly distinguished, we use the fused representation z_i^F as a reference and compare their similarities to it. The resulting similarity difference is used to compute the relation score s_{ij}^m, where a smaller difference indicates greater relation ambiguity. Meanwhile, evidential fusion is used to estimate the sample reliabilities r_i and r_j, which are combined to obtain the pair reliability r_{ij}. RCL jointly uses s_{ij}^m and r_{ij} to weight each negative pair, reducing the influence of ambiguous relations and pairs involving unreliable samples.

The fused predictive distribution is finally used for clustering. We use DDC (Kampffmeyer et al., 2019) as the clustering objective, with its formulation provided in Appendix A.1. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{DDC}} + \lambda_1 \mathcal{L}_{\mathrm{CGIB}} + \lambda_2 \mathcal{L}_{\mathrm{RCL}}, \tag{1}$$

where λ_1 and λ_2 control the contributions of CGIB and RCL. The final cluster assignment is $\hat{y}_i = \arg\max_k p_{ik}^F$.

3.3 Cross-Guided Information Bottleneck

For a target modality m, we use the remaining modalities to distinguish information supported across modalities from modality-specific information in its representation. Let X^m and Z^m denote the random variables for modality m and its latent representation. We denote all supporting modalities by $X^{-m} = (X^1, \ldots, X^{m-1}, X^{m+1}, \ldots, X^M)$. Since Z^m is generated only from X^m, the encoder satisfies $q_m(z \mid x^m, x^{-m}) = q_m(z \mid x^m)$. Thus, Z^m is conditionally independent of X^{-m} given X^m, and $I(Z^m; X^{-m} \mid X^m) = 0$. By the chain rule of mutual information,

$$I(Z^m; X^m) = I(Z^m; X^m \mid X^{-m}) + I(Z^m; X^{-m}). \tag{2}$$

Assuming that information useful for clustering is largely shared across modalities, $I(Z^m; X^m \mid X^{-m})$ characterizes information retained from X^m that is specific to the target modality, whereas $I(Z^m; X^{-m})$ characterizes information in Z^m supported by the other modalities. We define the cross-guided information bottleneck objective as

$$\mathcal{J}_{\mathrm{CGIB}} = \frac{1}{M} \sum_{m=1}^{M} \left[I(Z^m; X^{-m}) - I(Z^m; X^m \mid X^{-m}) \right]. \tag{3}$$

The objective favors information supported by the other modalities while suppressing modality-specific information. Next, we construct a cross-modal guidance distribution from the supporting modalities for each target modality. Each modality-specific latent distribution is parameterized as $q_m(z \mid x_i^m) = \mathcal{N}(\mu_i^m, \mathrm{diag}(v_i^m))$, following the diagonal Gaussian formulation in (Alemi et al., 2017).

The usefulness of a supporting modality varies across samples. To construct the guidance for target modality m, we weight the supporting modalities separately for each sample. For a batch of B samples, we form a reference representation for each sample i from the latent means of its supporting modalities: $\eta_i^{-m} = \frac{1}{M-1} \sum_{k \neq m} \mu_i^k$. For each supporting modality $n \neq m$, we measure its relation to this reference using cosine similarity

$$c_{ij}^{m,n} = \cos\left(\eta_i^{-m}, \mu_j^n\right). \tag{4}$$

The matched relation $c_{ii}^{m,n}$ compares the reference of sample i with the representation of the same sample i in modality n. A mismatched relation $c_{ij}^{m,n}$, $j \neq i$, instead compares this reference with the representation of another sample j in modality n. A supporting modality is more useful for sample i when its matched relation is clearly separated from the mismatched relations. We quantify this separation using their similarity gap:

$$a_{ij}^{m,n} = 1 - \exp\left(-\left|c_{ii}^{m,n} - c_{ij}^{m,n}\right|\right), \qquad j \neq i. \tag{5}$$

A larger $a_{ij}^{m,n}$ indicates a clearer distinction between the matched and mismatched relations. We average these distinction scores over mismatched samples as $\gamma_{in}^{m} = \frac{1}{B-1}\sum_{j\neq i} a_{ij}^{m,n}$. We then apply softmax to $\{\gamma_{in}^{m}\}_{n\neq m}$ across the supporting modalities to obtain the sample-specific guidance weights $\{\omega_{in}^{m}\}_{n\neq m}$. A supporting modality that consistently distinguishes the matched relation from mismatched relations receives a larger weight.

Using these weights, we aggregate the latent distributions of the supporting modalities into a weighted mixture and approximate it by a diagonal Gaussian through moment matching. The resulting mean is $\mu_i^{-m} = \sum_{n\neq m} \omega_{in}^{m}\mu_i^{n}$, and the variance is

$$v_i^{-m} = \sum_{n\neq m} \omega_{in}^{m} v_i^{n} + \sum_{n\neq m} \omega_{in}^{m}\left(\mu_i^{n} - \mu_i^{-m}\right)^2. \tag{6}$$

The first term accounts for the variance within each supporting latent distribution, while the second accounts for the dispersion among their latent means. The derivation is provided in Appendix A.2. The cross-modal guidance distribution for modality m is

$$g_m(z \mid x_i^{-m}) = \mathcal{N}\left(\mu_i^{-m}, \mathrm{diag}(v_i^{-m})\right). \tag{7}$$

When only two modalities are available, each target modality has only one supporting modality, so its guidance distribution reduces to the latent distribution of that modality. With more modalities, the guidance distribution combines multiple supporting latent distributions according to their sample-specific guidance weights.

3.4 Reliable Contrastive Learning

RCL determines how strongly each negative relation is trusted by weighting the corresponding negative pair according to relation ambiguity and sample reliability. A negative relation is more ambiguous when its mismatched similarity is close to the corresponding matched similarity, while a sample is less reliable when its predictions are uncertain or inconsistent across modalities. We quantify these two factors separately and combine them to determine the weight of each negative pair.

Following TMC (Han et al., 2021), we use an evidential network to produce non-negative evidence $\mathbf{e}_i^m = [e_{i1}^m, \ldots, e_{iK}^m]$ for each modality. The corresponding Dirichlet parameters are $\alpha_{ik}^m = e_{ik}^m + 1$, with total strength $S_i^m = \sum_{k=1}^{K} \alpha_{ik}^m$. The predictive distribution is $\mathbf{p}_i^m = [p_{i1}^m, \ldots, p_{iK}^m]$, where $p_{ik}^m = \alpha_{ik}^m / S_i^m$, and the estimated uncertainty is $u_i^m = K/S_i^m$. The modality-specific opinions are fused using the Dempster–Shafer rule, as detailed in Appendix A.3. This yields the fused predictive distribution $\mathbf{p}_i^F = [p_{i1}^F, \ldots, p_{iK}^F]$ and fused uncertainty u_i^F.

We measure the average disagreement between the predictive distributions and the fused predictive distribution using total variation distance:

$$\delta_i = \frac{1}{M}\sum_{m=1}^{M} \frac{1}{2}\left\|\mathbf{p}_i^m - \mathbf{p}_i^F\right\|_1. \tag{8}$$

A smaller δ_i indicates stronger agreement across modalities. We define the sample reliability and the corresponding pair reliability as

$$r_i = (1 - u_i^F)(1 - \delta_i), \qquad r_{ij} = \sqrt{r_i r_j}. \tag{9}$$

A larger r_{ij} indicates that both samples have low uncertainty and strong agreement across modalities.

We next compute a relation score to characterize relation ambiguity. Specifically, we apply softmax to $\{-u_i^m\}_{m=1}^{M}$ across modalities to obtain the weights $\{\rho_i^m\}_{m=1}^{M}$, so that modalities with lower

uncertainty receive larger weights. The fused representation is $z_i^F = \sum_{m=1}^{M} \rho_i^m z_i^m$. For modality m, we measure its relation to the fused representation using cosine similarity, $c_{ij}^m = \cos(z_i^F, z_j^m)$. The matched relation c_{ii}^m compares the fused representation of sample i with the representation of the same sample i in modality m. A mismatched relation c_{ij}^m, $j \neq i$, compares it with the representation of another sample j in modality m.

Following the formulation in Eq. (5), we define

$$s_{ij}^m = 1 - \exp\left(-\left|c_{ii}^m - c_{ij}^m\right|\right), \qquad j \neq i. \tag{10}$$

A larger s_{ij}^m indicates a clearer distinction between the matched and mismatched relations, while a smaller value indicates greater relation ambiguity. The negative weight is $w_{ij}^m = s_{ij}^m r_{ij}$. A large w_{ij}^m typically corresponds to a well-separated negative pair involving reliable samples.

For a batch of B samples, the RCL loss is

$$\mathcal{L}_{\mathrm{RCL}} = -\frac{1}{MB} \sum_{m=1}^{M} \sum_{i=1}^{B} \log \frac{\exp(c_{ii}^m/\tau)}{\exp(c_{ii}^m/\tau) + \sum_{j \neq i} w_{ij}^m \exp(c_{ij}^m/\tau)}, \tag{11}$$

where τ is the temperature parameter.

3.5 Optimization

We optimize the two information terms in Eq. (3) separately. For $I(Z^m; X^m \mid X^{-m})$, introducing the guidance distribution $g_m(z \mid x^{-m})$ gives the following variational upper bound:

$$I(Z^m; X^m \mid X^{-m}) \leq \mathbb{E}_{p(x^m, x^{-m})}\left[D_{\mathrm{KL}}\left(q_m(z \mid x^m) \| g_m(z \mid x^{-m})\right)\right]. \tag{12}$$

Since $q_m(z \mid x_i^m)$ and $g_m(z \mid x_i^{-m})$ are diagonal Gaussians, the KL divergence has a closed form. We define the KL loss for modality m as

$$\mathcal{L}_{\mathrm{KL}}^m = \frac{1}{Bd} \sum_{i=1}^{B} D_{\mathrm{KL}}\left(q_m(z \mid x_i^m) \| g_m(z \mid x_i^{-m})\right), \tag{13}$$

where d is the latent dimension. The derivation of the upper bound and the closed-form KL divergence are provided in Appendix A.4. For the preservation term $I(Z^m; X^{-m})$, we use a dependency regularizer $\mathcal{R}_m$ inspired by the mutual information formulation in IIC (Ji et al., 2019), as detailed in Appendix A.5. The regularizer measures the dependence between the modality-specific latent mean and its cross-modal guidance mean, with a larger value indicating stronger dependence. Combining it with the KL upper bound gives the trainable CGIB loss

$$\mathcal{L}_{\mathrm{CGIB}} = \frac{1}{M} \sum_{m=1}^{M} \left(\mathcal{L}_{\mathrm{KL}}^m - \mathcal{R}_m\right). \tag{14}$$

The KL term limits modality-specific information, while $\mathcal{R}_m$ encourages dependence between the modality-specific representation and its cross-modal guidance.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on five multi-modal datasets: Caltech-2V, Caltech-3V, WVU, MIRFlickr, and Prokaryotic. Caltech-2V (Fei-Fei et al., 2004) contains 1,400 images from 7 object categories, represented by two feature modalities: Wavelet Moments (Shen & Ip, 1999) and CENTRIST (Wu & Rehg, 2011). Caltech-3V uses the same images and categories, with LBP features (Ojala et al., 2002) added as the third modality. WVU (Ramagiri et al., 2011) is a human action dataset with four modalities, where video data are represented using features extracted by the Harris3-D detector and spatiotemporal interest points with HoG/HoF descriptors. MIR-Flickr (Huiskes & Lew, 2008) contains 12,154 image-text pairs from 6 categories, represented by visual and textual features. Prokaryotic (Brbic & Kopriva, 2018) consists of 551 prokaryotic species represented by three heterogeneous modalities: textual annotations, proteome composition profiles, and gene repertoire representations.

Table 1: Clustering performance ACC and NMI (%) on five multi-modal datasets. The **bold** denotes the best results and underline denotes the second-best results.

Method	Caltech-2V		Caltech-3V		WVU		MIRFlickr		Prokaryotic	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KM	41.6	30.5	46.3	31.3	30.8	37.2	40.9	22.5	54.5	19.8
Ncuts (TPAMI'00)	39.9	31.2	42.6	25.4	55.9	41.9	48.4	26.1	52.6	22.5
AmKM	44.6	35.2	46.9	31.5	27.9	25.1	41.0	21.6	54.6	22.9
AmNcuts (TPAMI'00)	42.8	25.2	43.7	25.5	58.3	55.0	48.2	26.2	53.0	28.1
SiMVC (CVPR'21)	50.8	47.1	56.9	50.4	46.6	45.2	45.6	26.3	50.1	19.6
CoMVC (CVPR'21)	46.6	42.6	54.1	50.4	42.3	44.4	49.3	30.6	51.0	21.6
MFLVC (CVPR'22)	60.6	52.8	63.1	56.6	58.2	51.3	53.8	32.8	57.5	23.9
DealMVC (ACM MM'23)	60.0	50.0	59.5	56.8	55.2	56.4	49.3	32.1	53.4	29.4
DIVIDE (AAAI'24)	64.1	52.9	67.8	56.2	49.9	50.0	52.3	33.5	55.9	31.3
ICMVC (AAAI'24)	49.6	37.9	64.7	53.7	38.3	39.0	43.5	24.4	52.3	24.5
MSCIB (KBS'24)	63.5	53.9	72.1	61.7	55.7	56.0	49.6	32.1	52.1	26.3
SDCIB (ICML'25)	67.5	59.2	69.0	62.4	59.2	60.0	50.4	32.9	53.0	30.0
PDMC-RCL (TIP'25)	62.5	52.4	69.7	58.4	58.0	49.1	52.7	33.0	56.3	27.7
MCSF (CVPR'26)	57.4	49.4	72.1	61.4	52.3	63.3	49.8	32.7	54.4	25.9
RAV (ICLR'26)	64.9	53.1	74.6	64.8	56.3	60.4	51.3	33.0	59.5	29.3
Ours	**73.3**	**63.9**	**79.1**	**71.6**	**64.5**	**67.2**	**59.4**	**37.0**	**67.3**	**41.6**

Implementation Details. All experiments are conducted using Python 3.8.20 and PyTorch 1.13.1 with CUDA 11.7 on an NVIDIA RTX 4090D GPU. The model is optimized with Adam using an initial learning rate of 0.001 and is trained for 100 epochs. The latent representation dimension is set to 256. The hyperparameters λ_1 and λ_2 are selected by grid search within $[10^{-4}, 1]$. The Gaussian kernel parameter in DDC is fixed to 0.15, while the temperature parameter τ is fixed to 0.1. We compare CGIB-RCL with four classical clustering baselines and eleven recent MMC methods. The classical baselines include K-Means (KM), Normalized Cuts (Ncuts) (Shi & Malik, 2000), and their all-modal versions. The recent methods include SiMVC and CoMVC (Trosten et al., 2021), MFLVC (Xu et al., 2022), DealMVC (Yang et al., 2023), DIVIDE (Lu et al., 2024), ICMVC (Chao et al., 2024), MSCIB (Yan et al., 2024), SDCIB (Lou et al., 2025c), PDMC-RCL (Lou et al., 2025a), MCSF (Cai et al., 2026), and RAV (Peng et al., 2026). For all compared methods, we follow their official implementations and use the recommended parameters or tuned configurations that yield their best performance.

4.2 Experimental Results

Results Analysis. As shown in Table 1, CGIB-RCL achieves the best ACC and NMI on all five datasets. On Prokaryotic, it reaches 67.3% ACC and 41.6% NMI, improving over the second-best results by 7.8% and 10.3%, respectively. On Caltech-3V, CGIB-RCL obtains 79.1% ACC and 71.6% NMI, surpassing RAV by 4.5% in ACC and 6.8% in NMI. These results demonstrate the consistent effectiveness of CGIB-RCL across the evaluated datasets.

Compared with recent MMC methods that combine IB with contrastive learning, including MSCIB and SDCIB, CGIB-RCL also performs better across all five datasets. On Caltech-3V, it improves over MSCIB by 7.0% in ACC and 9.9% in NMI, while the gains over SDCIB are 10.1% and 9.2%, respectively. These improvements support the benefit of explicitly regulating both information preservation and negative relations during learning. Specifically, CGIB adapts the contributions of supporting modalities for sample-specific information preservation, while RCL reduces the influence of ambiguous negative relations and pairs involving unreliable samples.

Parameter Analysis. We vary both λ_1 and λ_2 from 0.1 to 1.0 in steps of 0.1, as shown in Figure 2. Performance changes modestly across most settings. On Caltech-2V and Caltech-3V, ACC and NMI tend to decrease as λ_1 increases, while the other datasets are less sensitive to this parameter. In

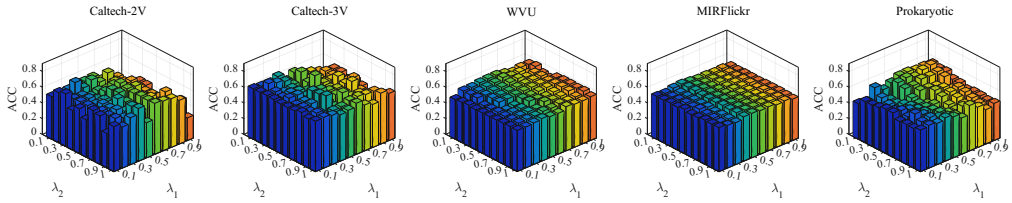

Figure 2: Parameter analysis of λ_1 and λ_2 on five datasets.

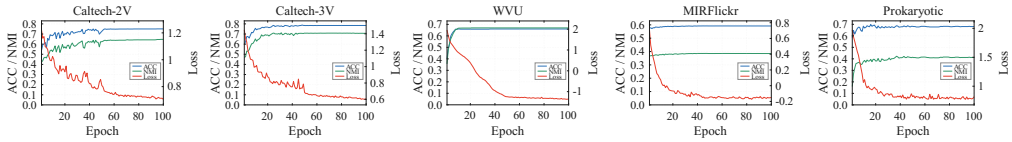

Figure 3: Convergence analysis on five datasets.

comparison, performance is generally less sensitive to λ_2, except for a gradual decrease on WVU at larger values. Overall, CGIB-RCL is relatively insensitive to the exact choices of λ_1 and λ_2.

Convergence Analysis. Figure 3 shows the overall loss, ACC, and NMI during training. On all five datasets, the loss decreases rapidly during the early epochs and then levels off, while ACC and NMI improve during training. Although both metrics fluctuate slightly at the beginning, they largely stabilize after about 60 epochs. No substantial change in clustering performance is observed during the later stage of training, suggesting that the model has largely converged within 100 epochs.

Visualization Analysis. Figure 4 shows the learned representations at different training epochs using t-SNE. Samples overlap considerably at the beginning of training, while the clusters become more compact and better separated over time. This trend is consistent with the improvement in clustering performance during training.

4.3 Ablation Study

Effect of the Main Components. Table 2 evaluates CGIB and RCL separately and in combination. CGIB improves both ACC and NMI over DDC on four of the five datasets, with particularly clear improvements on WVU and Prokaryotic. RCL also improves clustering performance on most datasets and shows its largest improvement on Caltech-3V.

The effect of either component alone varies across datasets, and neither consistently improves performance across all datasets. In contrast, combining CGIB and RCL achieves the best ACC and NMI on all five datasets. This pattern supports their complementary roles, with CGIB regulating information preservation and RCL regulating the influence of negative pairs.

Effect of the Design Choices. Table 3 further evaluates several key design choices. We first include the target modality in the construction of its own cross-modal guidance (w/ Target Modality in Guidance). This modification reduces both ACC and NMI on all five datasets, with particularly

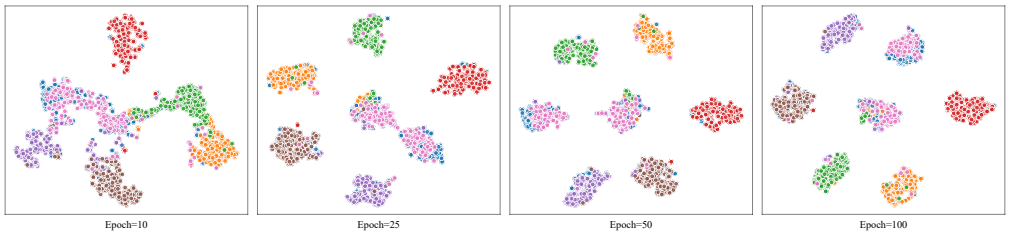

Figure 4: t-SNE visualization of the learned representations at different training epochs.

Table 2: Ablation study of the main components of the proposed method on five datasets.

Setting	Caltech-2V		Caltech-3V		WVU		MIRFlickr		Prokaryotic	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
DDC	66.8	56.5	64.7	55.7	56.9	56.4	54.4	33.4	62.4	32.9
DDC + CGIB	66.5	57.1	66.6	58.4	60.5	62.1	58.1	36.4	66.4	39.1
DDC + RCL	66.5	56.0	69.4	63.3	55.7	54.4	56.5	35.2	65.3	37.6
Ours	**73.3**	**63.9**	**79.1**	**71.6**	**64.5**	**67.2**	**59.4**	**37.0**	**67.3**	**41.6**

Table 3: Ablation study of the key design choices on five datasets.

Setting	Caltech-2V		Caltech-3V		WVU		MIRFlickr		Prokaryotic	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
w/ Target Modality in Guidance	63.9	54.9	72.2	61.3	58.8	56.5	57.5	36.3	65.2	38.7
w/o Relation Score	63.8	55.7	67.1	60.8	63.1	61.1	57.2	35.9	65.3	38.9
w/o Pair Reliability	65.9	55.4	64.1	59.1	57.2	60.2	57.9	36.9	66.1	39.7
w/o Negative Weighting	62.9	55.5	64.6	57.2	51.7	55.6	53.0	35.2	63.3	37.5
Ours	**73.3**	**63.9**	**79.1**	**71.6**	**64.5**	**67.2**	**59.4**	**37.0**	**67.3**	**41.6**

clear degradation on Caltech-3V and WVU. This supports constructing the guidance only from the supporting modalities, since including the target modality may allow modality-specific information to enter the guidance used to regulate that same modality.

We next examine the two factors in RCL. Removing the relation score makes the negative weights depend only on pair reliability, while removing pair reliability leaves only the relation score. Both variants perform worse than the full model on all five datasets. When both factors are removed and all negative pairs receive equal weights, the performance degradation becomes larger on several datasets. These comparisons support modeling relation ambiguity and sample reliability separately, as the two factors provide complementary signals for determining the contribution of negative pairs.

4.4 Mechanism Analysis

We further analyze the main mechanisms of CGIB-RCL, with details provided in Appendix A.6. First, to examine CGIB under degraded supporting information, we corrupt one supporting modality and examine its guidance weight and resulting guidance. As corruption increases, the corrupted modality receives lower guidance weights, and the guidance is less affected. This shows that CGIB reduces the contribution of corrupted supporting information when constructing cross-modal guidance. Second, to examine the two factors used by RCL, we compare the relation scores of false and true negatives and analyze sample reliability under corruption. False negatives receive lower relation scores than true negatives, while corrupted samples receive lower reliability. Together, these results show that RCL reduces the influence of ambiguous negative relations and pairs involving unreliable samples.

5 Conclusion

We propose CGIB-RCL, a novel reliable multi-modal framework that jointly addresses what information to preserve and how strongly negative relations should be trusted during learning. CGIB constructs sample-specific guidance from the supporting modalities to preserve information supported across modalities while reducing modality-specific content. RCL weights negative pairs according to relation ambiguity and sample reliability, reducing the influence of ambiguous relations and pairs involving unreliable samples. Experiments on five datasets demonstrate consistent improvements over the compared methods, with ablation and mechanism analyses supporting the proposed designs. Future work will explore extending CGIB-RCL to incomplete multi-modal data and developing more flexible strategies when the number of clusters is unknown in advance.

AI USE STATEMENT

In accordance with academic integrity guidelines, we confirm that no Large Language Models (LLMs) were used in the development of ideas, experimental work, analysis, or writing of the paper. All relevant material was developed and written independently by the authors.

ETHICS STATEMENT

This work does not raise any specific ethical concerns.

REPRODUCIBILITY STATEMENT

The main paper and appendix provide the details needed to reproduce our method and experiments.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Antonio Almudévar, José Miguel Hernández-Lobato, Sameer Khurana, Ricard Marxer, and Alfonso Ortega. Aligning multimodal representations through an information bottleneck. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Maria Brbic and Ivica Kopriva. Multi-view low-rank sparse subspace clustering. *Pattern Recognit.*, 73:247–258, 2018.

Bing Cai, Xiaoli Wang, Gui-Fu Lu, and Zechao Li. Multi-hierarchical contrastive spectral fusion for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39617–39626, June 2026.

Guoqing Chao, Yi Jiang, and Dianhui Chu. Incomplete contrastive multi-view clustering with high-confidence guiding. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 11221–11229. AAAI Press, 2024.

Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 16706–16715. IEEE, 2023.

Chenhang Cui, Yazhou Ren, Jingyu Pu, Jiawei Li, Xiaorong Pu, Tianyi Wu, Yutao Shi, and Lifang He. A novel approach for effective multi-view clustering with information-theoretic perspective. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2004, Washington, DC, USA, June 27 - July 2, 2004*, pp. 178. IEEE Computer Society, 2004.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Jie Hu, Chenghao Yang, Kai Huang, Hongjun Wang, Bo Peng, and Tianrui Li. Information bottleneck fusion for deep multi-view clustering. *Knowl. Based Syst.*, 289:111551, 2024.

Shizhe Hu, Binyan Tian, Weibo Liu, and Yangdong Ye. Self-supervised trusted contrastive multi-view clustering with uncertainty refined. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 17305–17313. AAAI Press, 2025a.

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, 2025b.

Mark J. Huiskes and Michael S. Lew. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, MIR 2008, Vancouver, British Columbia, Canada, October 30-31, 2008*, pp. 39–43. ACM, 2008.

Xu Ji, Andrea Vedaldi, and João F. Henriques. Invariant information clustering for unsupervised image classification and segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 9864–9873. IEEE, 2019.

Michael Kampffmeyer, Sigurd Løkse, Filippo Maria Bianchi, Lorenzo Livi, Arnt-Børre Salberg, and Robert Jenssen. Deep divergence-based approach to clustering. *Neural Networks*, 113:91–101, 2019.

Tianchao Li, Shujian Yu, Xinrui Zu, Zhaolong Wei, Jeremy Gummeson, Jack C. P. Cheng, and Robert Jenssen. OVA-IB: one vs all information bottleneck for multi-modal alignment. *CoRR*, abs/2605.29900, 2026. URL https://doi.org/10.48550/arXiv.2605.29900.

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, 2025a.

Zhengzheng Lou, Hang Xue, Chaoyang Zhang, and Shizhe Hu. A peer-review look on multi-modal clustering: An information bottleneck realization method. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025b.

Zhengzheng Lou, Ke Zhang, Yucong Wu, and Shizhe Hu. Super deep contrastive information bottleneck for multi-modal clustering. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025c.

Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. Decoupled contrastive multi-view clustering with high-order random walks. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pp. 14193–14201. AAAI Press, 2024.

Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7): 971–987, 2002.

Shuangmei Peng, Zhe Chen, Tianyang Xu, and Xiaojun Wu. Relationship alignment for view-aware multi-view clustering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=uRA9cT4MK6.

Sricharan Ramagiri, Rahul Kavi, and Vinod Kulathumani. Real-time multi-view human action recognition using a wireless camera network. In *2011 Fifth ACM/IEEE International Conference on Distributed Smart Cameras, Ghent, Belgium, Aug. 22-25, 2011*, pp. 1–6. IEEE, 2011.

Yazhou Ren, Zichen Wen, Junlong Ke, Chenhang Cui, Yonghao Huang, Xinyue Chen, Philip S. Yu, and Lifang He. Enhancing multi-view clustering: A sufficient information-theoretic approach for consistency acquisition and redundancy elimination. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48 (7):7420–7435, 2026.

Dinggang Shen and Horace Ho-Shing Ip. Discriminative wavelet shape descriptors for recognition of 2-d patterns. *Pattern Recognit.*, 32(2):151–165, 1999.

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

Naftali Tishby, Fernando C. N. Pereira, and William Bialek. The information bottleneck method. *CoRR*, physics/0004057, 2000.

Daniel J. Trosten, Sigurd Løkse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 1255–1265. Computer Vision Foundation / IEEE, 2021.

Daniel J. Trosten, Sigurd Løkse, Robert Jenssen, and Michael C. Kampffmeyer. On the effects of self-supervision and contrastive alignment in deep multi-view clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 23976–23985. IEEE, 2023.

Jing Wang and Songhe Feng. Contrastive and view-interaction structure learning for multi-view clustering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pp. 5055–5063. ijcai.org, 2024.

Jianxin Wu and James M. Rehg. CENTRIST: A visual descriptor for scene categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1489–1501, 2011.

Qilong Wu, Yiyang Shao, Jun Wang, and Xiaobo Sun. Learning optimal multimodal information bottleneck representations. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 16030–16039. IEEE, 2022.

Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, and Xiaofeng Zhu. Investigating and mitigating the side effects of noisy views for self-supervised clustering algorithms in practical multi-view scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 22957–22966. IEEE, 2024a.

Lingling Xu, Haoran Xie, Fu Lee Wang, Xiaohui Tao, Weiming Wang, and Qing Li. Contrastive sentence representation learning with adaptive false negative cancellation. *Inf. Fusion*, 102:102065, 2024b.

Wenbiao Yan, Yiyang Zhou, Yifei Wang, Qinghai Zheng, and Jihua Zhu. Multi-view semantic consistency based information bottleneck for clustering. *Knowl. Based Syst.*, 288:111448, 2024.

Xihong Yang, Jiaqi Jin, Siwei Wang, Ke Liang, Yue Liu, Yi Wen, Suyuan Liu, Sihang Zhou, Xinwang Liu, and En Zhu. Dealmvc: Dual contrastive calibration for multi-view clustering. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pp. 337–346. ACM, 2023.

Xihong Yang, Siwei Wang, Fangdi Wang, Jiaqi Jin, Suyuan Liu, Yue Liu, En Zhu, Xinwang Liu, and Yueming Jin. Automatically identify and rectify: Robust deep contrastive multi-view clustering in noisy scenarios. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

A APPENDIX

In the supplemental material:

- Appendix A.1: Clustering Objective.
- Appendix A.2: Guidance Distribution.
- Appendix A.3: Evidential Fusion.
- Appendix A.4: Variational Upper Bound.
- Appendix A.5: Cross-Modal Dependency Regularizer.
- Appendix A.6: Mechanism Analysis.

A.1 CLUSTERING OBJECTIVE

Following DDC (Kampffmeyer et al., 2019), we construct the clustering objective from the fused predictive distributions. For a batch of B samples, let $\mathbf{G} = [\mathbf{p}_1^F; \ldots; \mathbf{p}_B^F] \in \mathbb{R}^{B\times K}$, and denote its k-th column by $\boldsymbol{\delta}_k$. The Gaussian kernel matrix $\mathbf{K}$ is constructed from the modality-averaged hidden representations $\mathbf{f}_i = \frac{1}{M}\sum_{m=1}^{M} \mathbf{f}_i^m$, where $\mathbf{f}_i^m$ denotes the hidden representation produced by the evidential network for modality m.

The first term is based on the Cauchy–Schwarz divergence and encourages separation between different cluster assignments:

$$\mathcal{L}_1 = \frac{1}{K}\sum_{k=1}^{K-1}\sum_{l>k} \frac{\boldsymbol{\delta}_k^\top \mathbf{K}\boldsymbol{\delta}_l}{\sqrt{\left(\boldsymbol{\delta}_k^\top \mathbf{K}\boldsymbol{\delta}_k\right)\left(\boldsymbol{\delta}_l^\top \mathbf{K}\boldsymbol{\delta}_l\right)}}. \tag{15}$$

The second term encourages the clustering predictions to approach different corners of the probability simplex. Let $\mathbf{c}_k \in \mathbb{R}^K$ denote the k-th canonical basis vector and define $A_{ik} = \exp(-\|\mathbf{p}_i^F - \mathbf{c}_k\|_2^2)$. Denoting the k-th column of $\mathbf{A}$ by $\mathbf{a}_k$, the second term is

$$\mathcal{L}_2 = \frac{1}{K}\sum_{k=1}^{K-1}\sum_{l>k} \frac{\mathbf{a}_k^\top \mathbf{K}\mathbf{a}_l}{\sqrt{\left(\mathbf{a}_k^\top \mathbf{K}\mathbf{a}_k\right)\left(\mathbf{a}_l^\top \mathbf{K}\mathbf{a}_l\right)}}. \tag{16}$$

The third term encourages orthogonality between different clustering outputs:

$$\mathcal{L}_3 = \mathrm{triu}(\mathbf{G}^\top \mathbf{G}), \tag{17}$$

where $\mathrm{triu}(\cdot)$ denotes the sum of the off-diagonal elements in the upper triangular part of the matrix. The final clustering objective is

$$\mathcal{L}_{\mathrm{DDC}} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3. \tag{18}$$

A.2 GUIDANCE DISTRIBUTION

For sample i and target modality m, consider the weighted mixture of the latent distributions from its supporting modalities:

$$\widetilde{g}_m(z \mid x_i^{-m}) = \sum_{n\neq m} \omega_{in}^m q_n(z \mid x_i^n), \qquad q_n(z \mid x_i^n) = \mathcal{N}\left(\mu_i^n, \mathrm{diag}(v_i^n)\right).$$

We approximate this mixture by a diagonal Gaussian through moment matching. Its first moment is

$$\mathbb{E}[Z] = \sum_{n\neq m} \omega_{in}^m \mu_i^n = \mu_i^{-m}.$$

For each component distribution, the element-wise second moment is

$$\mathbb{E}[Z^2 \mid n] = v_i^n + (\mu_i^n)^2.$$

The second moment of the mixture is therefore

$$\mathbb{E}[Z^2] = \sum_{n \neq m} \omega_{in}^m \left[v_i^n + (\mu_i^n)^2 \right].$$

Using $\mathrm{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$, we obtain

$$v_i^{-m} = \sum_{n \neq m} \omega_{in}^m \left[v_i^n + (\mu_i^n)^2 \right] - (\mu_i^{-m})^2.$$

Equivalently,

$$v_i^{-m} = \sum_{n \neq m} \omega_{in}^m v_i^n + \sum_{n \neq m} \omega_{in}^m \left(\mu_i^n - \mu_i^{-m} \right)^2,$$

which gives Eq. (6).

The conditional latent marginal $\bar{q}_m(z \mid x^{-m})$ is generally unavailable in closed form. We use the moment-matched Gaussian constructed from the supporting modalities as the auxiliary distribution in Eq. (12):

$$g_m(z \mid x_i^{-m}) = \mathcal{N}\left(\mu_i^{-m}, \mathrm{diag}(v_i^{-m}) \right).$$

A.3 Evidential Fusion

The belief mass corresponding to the modality-specific evidence defined in the main text is

$$b_{ik}^m = \frac{e_{ik}^m}{S_i^m}.$$

Each modality m is represented by the opinion

$$\mathcal{M}_i^m = (\{b_{ik}^m\}_{k=1}^K, u_i^m).$$

Given two modalities a and b, their opinions are combined using the Dempster–Shafer rule. Their conflict is

$$C_i = \sum_{k \neq l} b_{ik}^a b_{il}^b.$$

The fused belief mass and uncertainty are

$$b_{ik} = \frac{b_{ik}^a b_{ik}^b + b_{ik}^a u_i^b + b_{ik}^b u_i^a}{1 - C_i}, \qquad u_i = \frac{u_i^a u_i^b}{1 - C_i}.$$

The opinions from all M modalities are fused sequentially as

$$\mathcal{M}_i^F = \mathcal{M}_i^1 \oplus \mathcal{M}_i^2 \oplus \cdots \oplus \mathcal{M}_i^M = (\{b_{ik}^F\}_{k=1}^K, u_i^F).$$

The fused predictive probability is $p_{ik}^F = b_{ik}^F + u_i^F / K$. Since $u_i^F = K / S_i^F$, the corresponding fused strength is $S_i^F = K / u_i^F$, and the fused evidence is $e_{ik}^F = b_{ik}^F S_i^F = K b_{ik}^F / u_i^F$. We denote the fused evidence vector by $\mathbf{e}_i^F = [e_{i1}^F, \ldots, e_{iK}^F]$.

A.4 Variational Upper Bound

For modality m, the encoder induces the joint distribution

$$p(x^m, x^{-m}, z) = p(x^m, x^{-m}) q_m(z \mid x^m).$$

Since Z^m is generated only from X^m, $q_m(z \mid x^m, x^{-m}) = q_m(z \mid x^m)$, and therefore $I(Z^m; X^{-m} \mid X^m) = 0$.

Define the conditional latent marginal induced by the encoder as

$$\bar{q}_m(z \mid x^{-m}) = \int q_m(z \mid x^m) p(x^m \mid x^{-m}) \, dx^m.$$

The conditional mutual information can then be written as

$$I(Z^m; X^m \mid X^{-m}) = \mathbb{E}_{p(x^m, x^{-m})} \left[D_{\mathrm{KL}} \left(q_m(z \mid x^m) \| \bar{q}_m(z \mid x^{-m}) \right) \right].$$

For any auxiliary distribution $g_m(z \mid x^{-m})$,

$$\mathbb{E}_{p(x^m,x^{-m})}\left[D_{\mathrm{KL}}\left(q_m(z \mid x^m)\|g_m(z \mid x^{-m})\right)\right] = I(Z^m; X^m \mid X^{-m}) + \mathbb{E}_{p(x^{-m})}\left[D_{\mathrm{KL}}\left(\bar{q}_m(z \mid x^{-m})\|g_m(z \mid x^{-m})\right)\right].$$

Since the second term is non-negative,

$$I(Z^m; X^m \mid X^{-m}) \leq \mathbb{E}_{p(x^m,x^{-m})}\left[D_{\mathrm{KL}}\left(q_m(z \mid x^m)\|g_m(z \mid x^{-m})\right)\right].$$

The bound is tight when $g_m(z \mid x^{-m}) = \bar{q}_m(z \mid x^{-m})$. The identity also shows that the gap of the bound is the expected KL divergence between $\bar{q}_m(z \mid x^{-m})$ and $g_m(z \mid x^{-m})$. Hence, a guidance distribution closer to the conditional latent marginal gives a tighter bound.

In our formulation, $g_m(z \mid x_i^{-m})$ is the cross-modal guidance distribution defined in Section 3.3. Since both $q_m(z \mid x_i^m)$ and $g_m(z \mid x_i^{-m})$ are diagonal Gaussians, their KL divergence has the closed form

$$D_{\mathrm{KL}}\left(q_m(z \mid x_i^m)\|g_m(z \mid x_i^{-m})\right) = \frac{1}{2}\sum_{l=1}^{d}\left[\log\frac{v_{i,l}^{-m}}{v_{i,l}^m} + \frac{v_{i,l}^m + \left(\mu_{i,l}^m - \mu_{i,l}^{-m}\right)^2}{v_{i,l}^{-m}} - 1\right].$$

A.5 Cross-Modal Dependency Regularizer

We construct a dependency regularizer inspired by the mutual information formulation in IIC (Ji et al., 2019) to encourage dependence between the modality-specific latent mean and its cross-modal guidance mean.

We first map the latent means to non-negative representations so that the entries of the resulting co-occurrence matrix are non-negative before normalization:

$$h_i^m = \mathrm{softplus}(\mu_i^m), \qquad h_i^{-m} = \mathrm{softplus}(\mu_i^{-m}).$$

For a batch of B samples, we construct

$$\widetilde{P}_{ab}^m = \sum_{i=1}^{B} h_{i,a}^m h_{i,b}^{-m}, \qquad a, b \in \{1, \ldots, d\}.$$

We symmetrize $\widetilde{P}^m$ as

$$P^m = \frac{1}{2}\left(\widetilde{P}^m + (\widetilde{P}^m)^\top\right),$$

and normalize it as

$$P_{ab}^m \leftarrow \frac{P_{ab}^m}{\sum_{a=1}^{d}\sum_{b=1}^{d} P_{ab}^m}.$$

The corresponding marginals are

$$P_{a\cdot}^m = \sum_{b=1}^{d} P_{ab}^m, \qquad P_{\cdot b}^m = \sum_{a=1}^{d} P_{ab}^m.$$

The dependency regularizer is

$$\mathcal{R}_m = \sum_{a=1}^{d}\sum_{b=1}^{d} P_{ab}^m \log\frac{P_{ab}^m}{P_{a\cdot}^m P_{\cdot b}^m}.$$

A larger $\mathcal{R}_m$ indicates stronger dependence between the modality-specific latent mean and its cross-modal guidance mean. We use $\mathcal{R}_m$ as a dependency surrogate for the cross-modal information term $I(Z^m; X^{-m})$ in Eq. (3).

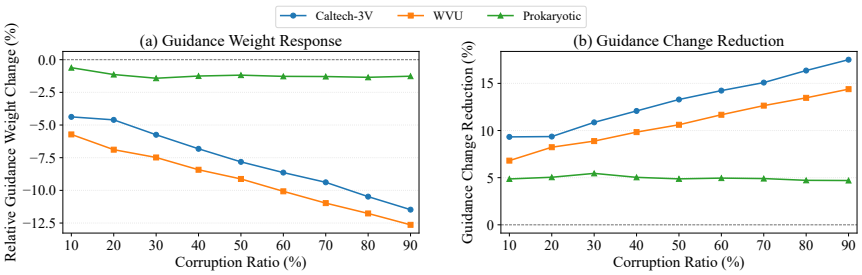

Figure 5: Controlled corruption analysis of CGIB. (a) Relative change in the guidance weight assigned to the corrupted modality. (b) Relative reduction in guidance change compared with uniform aggregation, measured by symmetric KL divergence.

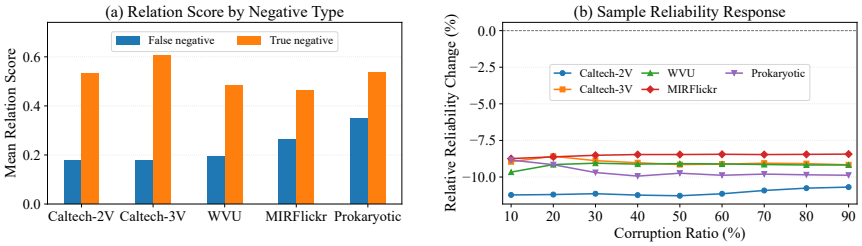

Figure 6: Analysis of RCL. (a) Mean relation scores of false and true negatives on the five datasets. (b) Relative change in sample reliability under corruption on the five datasets.

A.6 Mechanism Analysis

Guidance Weighting. We gradually corrupt one supporting modality to examine whether CGIB reduces its influence. For each dataset, one modality is used as the target modality and another as the corrupted supporting modality. We replace 10%–90% of the samples in the corrupted modality with Gaussian noise matched to the mean and standard deviation of each feature. Figure 5(a) reports the relative change $(\omega^{\text{corr}} - \omega^{\text{clean}})/\omega^{\text{clean}}$ in the guidance weight of the corrupted modality, where a negative value indicates a decrease in weight. The change is negative across all tested corruption ratios on Caltech-3V, WVU, and Prokaryotic.

Figure 5(b) compares the change in the guidance distribution under uniform aggregation and the proposed weighting. For each aggregation strategy, we compute the symmetric KL divergence between the clean and corrupted guidance distributions, denoted by D^{U}_{KL} and D^{W}_{KL}, respectively. We report the relative reduction $(D^{U}_{\text{KL}} - D^{W}_{\text{KL}})/D^{U}_{\text{KL}}$ as a percentage. The reduction is positive across all tested corruption ratios, showing that the proposed weighting makes the guidance distribution less affected by corruption than uniform aggregation.

Table 4: Analysis of reversed relations, where the mismatched similarity exceeds the matched similarity. Reversed Pairs denotes their fraction among all negative pairs, Mean Score reports the average relation score, and Score Mass denotes their fraction of the sum of all relation scores. Statistics are pooled over epochs 1, 10, 20, 50, and 100.

Dataset	Reversed (%)	Mean s (Rev.)	Mean s (Non-Rev.)	Rev. Score Mass (%)
Caltech-2V	2.02	0.014	0.429	0.07
Caltech-3V	3.01	0.047	0.523	0.28
WVU	2.01	0.005	0.389	0.03
MIRFlickr	7.42	0.015	0.326	0.38
Prokaryotic	9.14	0.041	0.366	1.13

Relation Score and Sample Reliability. We analyze the relation score and sample reliability separately in Figure 6. In Figure 6(a), ground-truth labels are used only for analysis to separate negative pairs into false negatives from the same class and true negatives from different classes. False negatives receive lower mean relation scores than true negatives on all five datasets. In Figure 6(b), one modality in each dataset is corrupted using the same Gaussian-noise procedure, and we report the relative change in reliability of the corrupted samples from the clean case. The change is negative across all tested corruption ratios on all five datasets, showing that corrupted samples receive lower reliability.

Effect of Reversed Similarity Relations. Since the relation score uses an absolute similarity gap, we also examine cases where the mismatched similarity exceeds the matched similarity, $c_{ij}^m > c_{ii}^m$. We refer to these cases as reversed relations. Reversed relations occur on all five datasets, but they account for only a small fraction of negative pairs and receive much lower relation scores than non-reversed pairs. Across the evaluated datasets, their contribution to the total relation score is at most 1.13%, suggesting that the use of the absolute similarity gap is only marginally affected by reversed relations in practice.