UNCERTAINTY-DRIVEN TRUSTWORTHY MULTI-MODAL INFORMATION BOTTLENECK CLUSTERING

Paper ID: 11848

ABSTRACT

The Information Bottleneck (IB) theory seeks to extract minimal sufficient representations and filter redundant noise, providing a solid foundation for multi-modal clustering. Although traditional IB-based methods demonstrate significant advantages, they still face several challenges. These include the inability to effectively model modality uncertainty, the tendency to yield overconfident incorrect predictions in the presence of modality conflicts, and the disconnection between representation learning and the final decision-making process. To address these issues, we propose a novel clustering method, the UNcertainty-drIven Trustworthy Multi-modal Information Bottleneck (UNIT-MIB). This approach adaptively regulates compression intensity based on uncertainty estimations and employs a dual-path fusion strategy. Specifically, it integrates semantic features via uncertainty-weighted aggregation and uncertainty-aware opinions via the Dempster-Shafer theory. Furthermore, based on Deng entropy, we formulate a Deng mutual information measure to characterize the discrepancy between semantic and evidential opinions, and introduce a stop-gradient-based directed alignment mechanism to impose asymmetric optimization. This mechanism steers semantic representation learning toward trustworthy evidential decision-making, thereby bridging the semantic-logical gap. Experimental results demonstrate that the proposed UNIT-MIB significantly outperforms state-of-the-art methods in terms of clustering accuracy across multiple real-world datasets.

1 INTRODUCTION

In real-world scenarios, multi-modal data are ubiquitous, with distinct modalities offering complementary information regarding the same entity. Multi-modal clustering seeks to harness this rich information to uncover latent common structures within the data in an unsupervised manner. Recently, Information Bottleneck (IB) theory (Tishby et al., 1999) has established a robust theoretical foundation for multi-modal clustering tasks. This theory is characterized by a principled capacity to extract minimal sufficient representations while effectively filtering redundant noise. The theoretical framework focuses on achieving an optimal balance between data compression and information preservation by creating a compressed representation T for an input X, which is optimized to retain features relevant to a target variable Y. The optimization process is governed by the following objective function:

$$\mathcal{L}_{min} = I(X;T) - \beta I(T;Y), \tag{1}$$

where $I(\cdot;\cdot)$ denotes mutual information, and the parameter β regulates the trade-off between the degree of compression and the relevance to the target. By minimizing $I(X;T)$ to discard irrelevant information while simultaneously maximizing $I(T;Y)$, the IB framework learns a compact yet informative representation of the input data (Alemi et al., 2017). Due to this elegant trade-off between information compression and preservation, IB has become a powerful tool in representation learning, finding wide applications across diverse communities such as natural language processing, computer vision, and data mining (Hu et al., 2024).

Inspired by these broad successes, researchers have extensively adapted IB theory for multi-modal clustering to learn discriminative representations. Initial efforts in this domain primarily focused on information disentanglement. For example, MIB (Federici et al., 2020) proposed a multi-modal IB method designed to identify non-shared information between two modalities. Similarly, MSCIB (Yan et al., 2024) utilizes a semantic consistency loss to guide the information bottleneck,

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.

using shared representations of multiple modalities to eliminate the private information of individual modalities. To further enhance clustering robustness, recent research has explored self-supervised weighting and guided reconstruction. PTIB (Lou et al., 2025) introduces a parameter-free "peer-review" mechanism to learn modality weights through mutual evaluation and trustworthiness scoring. SWIB (Lou et al., 2024) incorporates self-supervised information from pseudo-labels into the weighted IB framework, learning weights by combining modal quality with clustering consistency. Furthermore, MSDIB (Hu et al., 2025) proposes a "multi-aspect self-guided" strategy that comprehensively utilizes private information, shared information, and pseudo-labels through the information bottleneck. Beyond IB-based approaches, discriminative subspace clustering and graph-guided deep clustering have also been investigated to learn informative structures and discriminative representations from high-dimensional data (Huang et al., 2025; He et al., 2024).

Although IB-based multi-modal clustering methods (Yan et al., 2024; Hu et al., 2025) demonstrate significant advantages in representation learning, current approaches still face several critical challenges in complex multi-modal environments. First, existing IB methods generally lack the capacity to effectively model modality uncertainty. In practical applications, the reliability of distinct modalities varies significantly (Geng et al., 2021). Without quantifying these uncertainties, it is difficult to achieve robust adaptive regulation during feature extraction (Han et al., 2023). Second, when information conflicts arise among different modalities, current models tend to yield overconfident yet incorrect predictions (Jia et al., 2025). This issue is particularly severe in self-supervised clustering frameworks, where pseudo-labels derived from the model are inherently noisy. This low-quality guidance forces the model to learn erroneous representations, which subsequently produce increasingly flawed predictions. This process forms a vicious cycle, ultimately causing the confidence of the model to far exceed the actual accuracy. Finally, current clustering algorithms commonly suffer from a disconnection between representation learning and the final decision-making process. Specifically, a significant "semantic-logical gap" frequently exists between the semantic structures in the feature space and the logical reasoning at the decision level. This gap leads to clustering results that lack both consistency and reliability.

To overcome these challenges, we incorporate Evidential Deep Learning (EDL) and the Dempster-Shafer (D-S) theory to manage multi-modal uncertainty and conflicts. Unlike traditional Softmax, which often yields overconfident erroneous predictions, EDL quantifies epistemic uncertainty by modeling class probabilities via a Dirichlet distribution (Han et al., 2021). Consequently, the model can explicitly identify unreliable modalities. Furthermore, the D-S combination rule provides a principled mechanism for aggregating belief and uncertainty masses from multiple modalities while explicitly accounting for inter-modal conflict (Zhang et al., 2023).

Although evidence theory ensures trustworthiness at the decision level, its integration with IB theory at the representation learning level remains a non-trivial task. To bridge this gap, we leverage uncertainty to adaptively regulate IB compression intensity and introduce a directed guidance strategy. Unlike Shannon entropy, which does not explicitly represent the uncertainty mass in subjective opinions, Deng entropy (Deng, 2016) provides an information measure that accounts for both belief masses and epistemic uncertainty. Based on Deng entropy, we formulate a Deng mutual information measure to characterize the discrepancy between semantic representations and evidential opinions. To introduce directionality, we further employ a stop-gradient-based alignment strategy that guides semantic representations toward the trustworthy evidential consensus. In this way, the semantic-logical gap between representation learning and decision-making can be effectively reduced. Based on these principles, we propose UNIT-MIB, whose overall framework is illustrated in Figure 1.

Specifically, the UNIT-MIB method integrates information theory with evidential deep learning to construct a closed-loop trustworthy system from low-level feature representations to high-level decision reasoning. First, unlike traditional IB methods that employ a fixed compression intensity (Chalk et al., 2016), we utilize the epistemic uncertainty of each modality to dynamically regulate the trade-off coefficient β. This allows the model to adaptively strengthen compression to filter redundant noise for unreliable modalities, while preserving discriminative features for reliable ones. Second, we design a trustworthy dual-path fusion strategy. In the feature-level path, an uncertainty-weighted aggregation mechanism fuses unimodal features into a global semantic representation. Simultaneously, in the decision-level path, modality-specific subjective opinions are logically fused based on Dempster-Shafer theory to form a rigorous global joint opinion. Finally, we propose a directed alignment strategy based on Deng mutual information. Although the Deng mutual information measure is symmetric, directionality is introduced during optimization by stopping gradients through

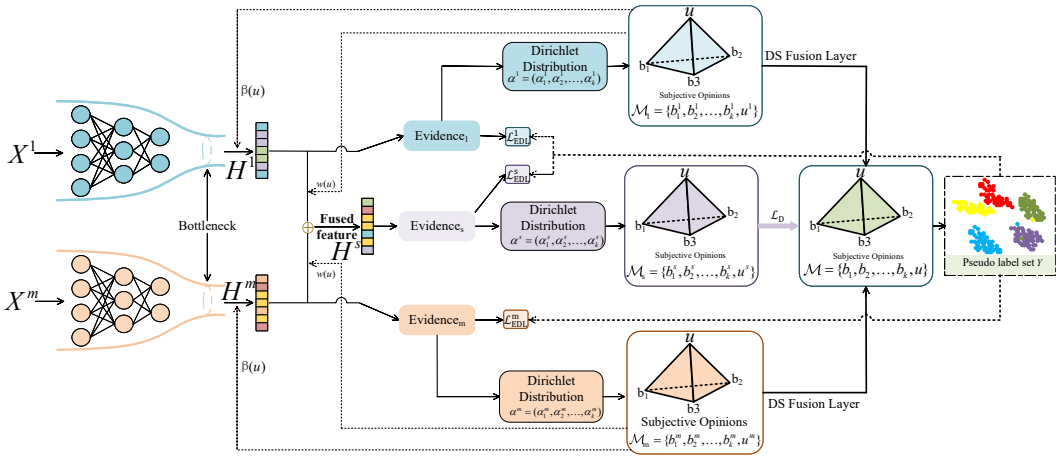

Figure 1: Illustration of the proposed UNIT-MIB framework. Modality-specific representations H^m are first learned through an uncertainty-driven information bottleneck with adaptive compression. A dual-path fusion strategy then constructs semantic and evidential consensus: the semantic path aggregates H^m into H^S using uncertainty-aware weighting $w(u)$, while the evidential path transforms modality-specific representations into subjective opinions $\mathcal{M}^m$ and recursively fuses them into the global consensus $\mathcal{M}^{DS}$ via Dempster-Shafer theory. A stop-gradient-based directed alignment loss $\mathcal{L}_D$ guides the feature-derived opinion toward the evidential consensus without reverse gradient propagation. The normalized belief masses of $\mathcal{M}^{DS}$ are further used to generate soft pseudo-labels Y for clustering supervision.

the evidential consensus. This mechanism guides the semantic representation toward trustworthy logical reasoning and facilitates a unified clustering consensus.

The primary innovations and contributions of this paper are summarized as follows:

- We propose a novel UNIT-MIB method that utilizes epistemic uncertainty as a regulatory factor to achieve modality-specific adaptive compression. This mechanism significantly enhances the capacity of the information bottleneck to autonomously balance compression levels, allowing for adaptive noise filtering and the preservation of purified features.
- We develop a trustworthy dual-path fusion architecture that harmonizes uncertainty-aware semantic fusion with Dempster-Shafer-based logical reasoning to establish a robust multi-modal consensus.
- We formulate a Deng mutual information measure to characterize the discrepancy between semantic and evidential opinions, and introduce a stop-gradient-based directed alignment mechanism. This design guides semantic representations toward trustworthy evidential consensus and reduces the semantic-logical gap between representation learning and decision-making.
- Evaluations across five benchmark datasets, including Caltech-2V, Scene15, and RGBD, demonstrate that UNIT-MIB significantly outperforms state-of-the-art methods in both clustering Accuracy (ACC) and Normalized Mutual Information (NMI).

2 The Proposed Method

2.1 Problem Formulation

In the multi-modal clustering task, we are given a dataset $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ containing N samples. Each sample x_i is composed of data from M different modalities, denoted as $x_i = \{x_i^1, x_i^2, \ldots, x_i^M\}$, where x_i^m represents the m-th modality for sample i, and C denotes the predefined number of clusters.

2.2 THE OVERALL FRAMEWORK

In this section, we present the proposed UNIT-MIB framework for trustworthy unsupervised multi-modal clustering. As illustrated in Figure 1, UNIT-MIB consists of four tightly coupled components: Uncertainty-driven Information Bottleneck, Uncertainty-weighted Feature Fusion, Trustworthy Decision-level Evidential Fusion, and Directed Alignment via Deng Mutual Information.

First, the **Uncertainty-driven Information Bottleneck** adaptively regulates the compression of each modality according to its epistemic uncertainty, encouraging the modality-specific representations H^m to preserve discriminative information while suppressing redundant noise. Second, the **Uncertainty-weighted Feature Fusion** aggregates these representations according to their reliability to construct the global semantic representation H^S. In parallel, the **Trustworthy Decision-level Evidential Fusion** transforms modality-specific representations into subjective opinions and recursively combines them through Dempster-Shafer theory to obtain the global evidential consensus $\mathcal{M}^{DS}$. Finally, the **Directed Alignment via Deng Mutual Information** aligns the feature-derived opinion $\mathcal{M}^S$ with $\mathcal{M}^{DS}$, where stop-gradient optimization makes the evidential consensus serve as a fixed guidance target for the semantic branch. Through these interconnected components, UNIT-MIB jointly performs uncertainty-aware representation learning, reliable multi-modal fusion, and semantic-evidential consensus alignment.

2.3 UNCERTAINTY-DRIVEN INFORMATION BOTTLENECK

To extract compact and discriminative features from each modality while suppressing modality-specific noise, we design the cross-modal alignment loss:

$$\mathcal{L}_{\mathrm{C}} = \sum_{m=1}^{M} I(X^m; H^m) - \sum_{m=1}^{M} \sum_{n=m+1}^{M} I(H^m; H^n), \tag{2}$$

where X^m is the input for modality m and H^m is its learned latent representation. The first term, $I(X^m; H^m)$, serves as a compression objective that encourages the network to discard irrelevant information from the input. The second term, $-I(H^m; H^n)$, aims to maximize the mutual information between representations of different modalities, forcing the unimodal encoders to learn a shared semantic space.

To further guide the representation learning with logically robust global evidence, we introduce the uncertainty-weighted evidential guidance objective. A core innovation of our framework is the integration of these objectives into an adaptive IB loss:

$$\mathcal{L}_{IB} = \mathcal{L}_{\mathrm{C}} - \sum_{m=1}^{M} \beta(u^m) I(H^m; \mathcal{M}^{DS}), \tag{3}$$

where $\mathcal{M}^{DS}$ represents the global fused opinion generated via the DS fusion layer, and $\beta(u^m)$ is a dynamic trade-off coefficient regulated by the epistemic uncertainty u^m of modality m.

Specifically, we formulate the dynamic trade-off coefficient $\beta(u^m)$ as a quadratic decay function of the epistemic uncertainty:

$$\beta(u^m) = (1 - u^m)^2. \tag{4}$$

We deliberately select this quadratic formulation over a standard linear decay $(1 - u^m)$ because it imposes a steeper and more aggressive penalty on modalities as the corresponding uncertainty increases. By maximizing the mutual information between the unimodal features H^m and the evidence-based consensus $\mathcal{M}^{DS}$, the encoder prioritizes the extraction of features that support the trustworthy global decision. When a modality exhibits high uncertainty, the quadratic decay rapidly diminishes the guidance strength $\beta(u^m)$ to near zero, thereby reducing the influence of potentially unreliable supervision. Conversely, for modalities with low uncertainty, a sufficiently large $\beta(u^m)$ is preserved to strengthen the alignment between the purified features and the trustworthy consensus.

2.4 UNCERTAINTY-WEIGHTED FEATURE FUSION

As the first path of the proposed trustworthy dual-path fusion strategy, the feature-level path aims to integrate the purified latent representations $\{H^1, H^2, \ldots, H^M\}$ into a unified global semantic

space H^S. Unlike traditional fusion methods that treat all modalities equally or use static weights, UNIT-MIB introduces a reliability-driven aggregation strategy that leverages the quantified epistemic uncertainty from the evidential branch to gate the information flow.

The global fused representation H^S is constructed as a dynamic weighted combination of unimodal features:

$$H^S = \sum_{m=1}^{M} w(u^m) \cdot H^m, \tag{5}$$

where $w(u^m)$ is the normalized reliability weight for the m-th modality. To ensure that the fusion process is dominated by high-confidence modalities while suppressed for noisy ones, we calculate $w(u^m)$ through an uncertainty-aware softmax-like formulation:

$$w(u^m) = \frac{\exp(1 - u^m)}{\sum_{n=1}^{M} \exp(1 - u^n)}. \tag{6}$$

In this formulation, $u^m \in [0, 1]$ represents the epistemic uncertainty derived from the subsequent evidential path. When a modality is corrupted by noise or exhibits low discriminability, the uncertainty u^m tends toward 1, leading to a minimal weight $w(u^m)$ to prevent error propagation. Conversely, a reliable modality with clear cluster structures yields a lower u^m, thereby contributing more significantly to the global representation H^S.

By adaptively adjusting the contribution of each modality in the latent space, this mechanism ensures that H^S preserves a high-fidelity semantic structure even in the presence of partial modality failure. This uncertainty-driven fusion serves as the semantic foundation for the subsequent directed alignment between the representation and evidence spaces.

2.5 Trustworthy Decision-level Evidential Fusion

Complementary to the feature-level path, this section details the second path of the proposed fusion strategy: the decision-level evidential reasoning. While Section 2.4 focuses on semantic aggregation, this path aims to construct a logically consistent consensus by transforming deterministic representations into probabilistic evidence. This establishes a rigorous framework for trustworthy reasoning through the explicit quantification of epistemic uncertainty.

2.5.1 Subjective Opinion Modeling

To quantify the uncertainty inherent in multi-modal data, we employ evidential deep learning to transform features into subjective opinions. For each path $v \in \mathcal{V} = \{1, \ldots, M, S\}$, a linear classifier with softplus activation extracts the evidence vector $e^v = [e_1^v, \ldots, e_C^v]^\top$. These values parameterize a Dirichlet distribution $\mathrm{Dir}(p^v|\alpha^v)$ with concentration parameters $\alpha_k^v = e_k^v + 1$. Based on these parameters, the belief mass for class k and the corresponding epistemic uncertainty are defined as $b_k^v = e_k^v/S^v$ and $u^v = C/S^v$, respectively, where $S^v = \sum_{k=1}^{C} \alpha_k^v$ represents the Dirichlet strength. This formulation yields a set of unimodal opinions $\{\mathcal{M}^1, \ldots, \mathcal{M}^M\}$ and a feature-derived opinion $\mathcal{M}^S$, where each opinion is defined as $\mathcal{M}^v = \{b_1^v, \ldots, b_C^v, u^v\}$. Notably, the uncertainty u^m derived here is utilized as the reliability weight in the feature fusion path described in Section 2.4.

2.5.2 Cross-branch Evidential Supervision

To ensure the discriminative power and calibration of each branch, we apply a multi-path evidential learning loss $\mathcal{L}_{EDL}$. For a dataset of N samples, the total loss over all samples i is formulated as:

$$\mathcal{L}_{EDL} = \sum_{i=1}^{N} \sum_{v \in \mathcal{V}} \Big(\|y_i - p_i^v\|_2^2 + \|p_i^v \odot (1 - p_i^v)\|_2^2 + \lambda_{KL} \mathcal{L}_{KL}^v \Big), \tag{7}$$

where $p_i^v = \alpha_i^v / S_i^v$ denotes the expected probability and y_i is the soft pseudo-label derived from the DS consensus. Specifically, the pseudo-label for class k is obtained by normalizing the fused belief masses as $y_{ik} = b_{ik}^{DS} / \sum_{c=1}^{C} b_{ic}^{DS}$. The pseudo-labels are initialized from an initial forward pass and refreshed at the end of each epoch using the current DS consensus, while remaining fixed within

each epoch. The first term minimizes the discrepancy between branch predictions and the consensus pseudo-labels, the second term penalizes predictive ambiguity, and the third term regularizes the Dirichlet evidence to prevent overconfident predictions.

2.5.3 Decision-level Fusion via DS Theory

In parallel to the semantic fusion, we implement an evidence-level fusion path that aggregates unimodal opinions into a joint mass. Dempster-Shafer (DS) theory provides the mathematical foundation to combine evidence from independent sources. For any two opinions $\mathcal{M}^1$ and $\mathcal{M}^2$, the joint belief $b_k^{1\oplus 2}$ is calculated as:

$$b_k^{1\oplus 2} = \frac{1}{1-K}\left(b_k^1 b_k^2 + b_k^1 u^2 + b_k^2 u^1\right). \tag{8}$$

The aggregated uncertainty is defined as $u^{1\oplus 2} = u^1 u^2/(1-K)$, where $K = \sum_{k\neq l} b_k^1 b_l^2$ denotes the degree of conflict between modalities. The final global joint opinion $\mathcal{M}^{DS}$ is obtained by recursively applying the combination operator, $\mathcal{M}^{DS} = \mathcal{M}^1 \oplus \mathcal{M}^2 \oplus \cdots \oplus \mathcal{M}^M$. This recursive fusion dynamically adjusts the consensus according to the uncertainty and conflict among modalities, yielding a robust and interpretable evidential consensus.

2.6 Directed Alignment via Deng Mutual Information

Although the feature-level path (Section 2.4) and the decision-level path (Section 2.5) provide complementary perspectives for multi-modal clustering, a potential semantic-logical gap may exist between the feature-fused opinion $\mathcal{M}^S$ and the evidence-fused consensus $\mathcal{M}^{DS}$. To enforce directed consistency and bridge this gap, we propose a directed alignment mechanism based on Deng mutual information.

Traditional Shannon entropy does not explicitly characterize the uncertainty mass u in subjective opinions. Therefore, we employ Deng entropy $H_D(\mathcal{M})$ to quantify the information content of an opinion $\mathcal{M} = \{b_1, \ldots, b_C, u\}$:

$$H_D(\mathcal{M}) = -\sum_{k=1}^{C} b_k \log_2(b_k) - u\log_2\left(\frac{u}{2^C-1}\right), \tag{9}$$

where C denotes the number of clusters. The second term explicitly incorporates the uncertainty mass and the corresponding power-set state space, allowing the information volume associated with epistemic uncertainty to be represented.

Based on Deng entropy, we formulate an information discrepancy between the feature-fused opinion $\mathcal{M}^S$ and the evidence-fused consensus $\mathcal{M}^{DS}$:

$$I_D(\mathcal{M}^S; \mathcal{M}^{DS}) = H_D\left(\frac{\mathcal{M}^S + \mathcal{M}^{DS}}{2}\right) - \frac{H_D(\mathcal{M}^S) + H_D(\mathcal{M}^{DS})}{2}. \tag{10}$$

The resulting measure characterizes the information discrepancy between the semantic and evidential opinions by jointly considering their belief and uncertainty components. A detailed derivation is provided in Appendix A.1.

Although $I_D(\mathcal{M}^S; \mathcal{M}^{DS})$ is symmetric with respect to its two arguments, the proposed alignment is made directional during optimization. Specifically, to prevent noisy semantic information from reversely affecting the logically robust evidential consensus, we apply a stop-gradient operation to $\mathcal{M}^{DS}$. The directed alignment loss is therefore defined as:

$$\mathcal{L}_D = I_D(\mathcal{M}^S; \mathrm{sg}(\mathcal{M}^{DS})), \tag{11}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Consequently, with respect to $\mathcal{L}_D$, gradients are propagated through the semantic branch while the evidential consensus serves as a fixed guidance target. By minimizing $\mathcal{L}_D$, the semantic representation is steered toward the logically robust consensus without reverse contamination, thereby reducing the semantic-logical gap and facilitating a unified and trustworthy clustering consensus.

2.7 OPTIMIZATION

To optimize the proposed UNIT-MIB framework, we integrate the uncertainty-driven information bottleneck, multi-path evidential learning, and directed alignment via Deng mutual information into a unified objective function and minimize it in an end-to-end manner:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{IB} + \alpha\mathcal{L}_{EDL} + \lambda\mathcal{L}_{D}, \tag{12}$$

where $\mathcal{L}_{IB}$ denotes the uncertainty-driven information bottleneck loss, $\mathcal{L}_{EDL}$ denotes the trustworthy evidential learning loss, and $\mathcal{L}_{D}$ denotes the directed alignment loss, with α and λ acting as trade-off parameters.

The mutual-information terms involved in $\mathcal{L}_{IB}$ are optimized through tractable variational bounds. Specifically, the information-compression term $I(X^m; H^m)$ is optimized using the VIB upper bound, while the cross-modal mutual information $I(H^m; H^n)$ is optimized using the InfoNCE lower bound. For the evidential guidance term $I(H^m; \mathcal{M}^{DS})$, we use the belief-mass component $b^{DS} \in \mathbb{R}^C$ to represent the DS consensus for contrastive guidance. It is first mapped into the latent representation space through a learnable projection head, $\tilde{z}^{DS} = g_\psi\big(\text{sg}(b^{DS})\big) \in \mathbb{R}^D$, and an InfoNCE objective is then applied between H^m and $\tilde{z}^{DS}$. Therefore, minimizing the VIB term suppresses modality-specific redundancy, whereas the contrastive objectives preserve shared cross-modal semantics and align modality-specific representations with the trustworthy evidential consensus. Detailed derivations are provided in Appendix A.2 and Appendix A.3.

The entire network is trained with the Adam optimizer, jointly refining the latent representation space and the evidential decision space so that the feature-level and decision-level paths co-evolve toward a unified, trustworthy clustering consensus.

3 EXPERIMENT

3.1 EXPERIMENTAL SETUP

Datasets. We conduct experiments on five widely used public multi-modal datasets, including Caltech-2V (Fei-Fei et al., 2004), RGBD (Li et al., 2009), Scene15 (Lazebnik et al., 2006), IAPR (Grubinger et al., 2006), and Flickr (Huiskes & Lew, 2008). Caltech-2V contains Wavelet moments (Shen & Ip, 1999) and CENTRIST features, RGBD consists of RGB and depth modalities, and Scene15 includes GIST (Oliva & Torralba, 2001), PHOG (Bosch et al., 2007), and LBP (Ojala et al., 2002) features. IAPR and Flickr are represented by visual and textual modalities. Detailed dataset descriptions and statistics are provided in Appendix A.4.

Implementation Details. UNIT-MIB is implemented using PyTorch 2.4.1 and Python 3.8. All experiments are conducted on a Windows 10 workstation equipped with an Intel Core i9 CPU and a 24GB NVIDIA GeForce RTX 4090D GPU. We train UNIT-MIB for 100 epochs and run each experiment independently 10 times with different random initializations. To mitigate the influence of local optima during training, we report the clustering results from the run with the minimum final training loss. Clustering performance is evaluated using two widely used metrics: Clustering Accuracy (ACC) and Normalized Mutual Information (NMI) (Strehl & Ghosh, 2002). Higher values of both metrics indicate better clustering performance.

For comparison, we select four traditional Single/All-modal clustering baselines and twelve representative deep multi-modal clustering methods. The traditional baselines include K-Means (KM), Normalized Cuts (Ncuts) (Shi & Malik, 2000), and their all-modal variants. For single-modal baselines, each modality is evaluated independently and the best result is reported. The deep multi-modal competitors include SiMVC and CoMVC (Trosten et al., 2021), MFLVC (Xu et al., 2022), DealMVC (Yang et al., 2023), ICMVC (Chao et al., 2024), MVCAN (Xu et al., 2024), CSOT (Zhang et al., 2024), ROLL (Sun et al., 2025), SCVT (Dong et al., 2025), COPER (Eisenberg et al., 2025), SPLMVC (Luo et al., 2026), and RAV (Peng et al., 2026).

3.2 EXPERIMENTAL RESULTS

Results Analysis. Table 1 reports the clustering performance of UNIT-MIB and the compared methods on five benchmark datasets. Overall, UNIT-MIB achieves the best ACC and NMI results on all

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

Methods	Caltech-2V		Scene15		RGBD		Flickr		IAPR	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KM	41.6	30.5	29.2	22.6	37.9	29.4	40.9	22.5	38.9	17.2
Ncuts	39.9	31.2	33.7	25.3	41.2	32.7	48.4	26.1	41.9	18.9
ALLKM	46.4	31.4	34.5	24.9	38.4	30.1	50.4	33.1	44.9	25.3
AllNcuts	42.8	25.2	32.8	26.7	42.5	31.9	48.2	26.2	42.2	18.9
SiMVC (2021)	51.1	36.9	32.8	29.3	38.6	35.6	45.6	26.3	42.7	18.5
CoMVC (2021)	59.2	49.2	37.6	31.3	40.3	38.7	49.3	30.6	46.7	21.5
MFLVC (2022)	61.5	53.6	29.8	27.9	43.2	30.8	51.8	32.8	47.3	22.6
DealMVC (2023)	47.6	37.9	35.7	33.6	29.7	26.8	49.3	32.1	35.0	10.8
ICMVC (2024)	49.6	37.9	36.2	34.2	31.8	25.7	43.5	24.4	37.1	16.8
MVCAN (2024)	46.5	37.5	37.9	41.4	21.9	17.9	50.5	31.5	34.4	17.2
CSOT (2024)	63.9	54.4	35.3	31.8	41.6	35.2	39.7	29.9	42.1	22.3
ROLL (2025)	45.7	35.8	31.1	28.6	39.5	37.3	46.2	27.3	38.2	17.4
SCVT (2025)	58.3	45.2	24.9	25.1	38.4	32.6	37.6	24.9	35.2	20.3
COPER (2025)	61.5	55.1	40.6	41.9	49.8	38.1	46.4	34.1	47.2	25.5
SPLMVC (2026)	60.8	48.7	36.2	32.9	40.5	33.9	42.7	30.8	47.6	20.9
RAV (2026)	62.3	54.6	41.2	40.7	42.3	33.5	40.8	30.9	43.8	23.5
UNIT-MIB (Ours)	**69.3**	**58.4**	**44.3**	**43.6**	**50.2**	**39.8**	**55.9**	**40.2**	**51.9**	**29.7**

datasets, consistently outperforming representative multi-modal clustering methods spanning 2021–2026. On Caltech-2V, UNIT-MIB achieves an ACC of 69.3% and an NMI of 58.4%, exceeding the second-best results of CSOT and COPER by 5.4% and 3.3%, respectively. On Scene15, UNIT-MIB improves the ACC from 41.2% achieved by RAV to 44.3%, while increasing the NMI from 41.9% achieved by COPER to 43.6%. For RGBD, UNIT-MIB achieves 50.2% ACC and 39.8% NMI, outperforming the second-best results by 0.4% and 1.1%, respectively. More substantial improvements are observed on Flickr, where UNIT-MIB achieves 55.9% ACC and 40.2% NMI, surpassing MFLVC and COPER by 4.1% and 6.1%, respectively. For IAPR, UNIT-MIB achieves 51.9% ACC and 29.7% NMI, improving upon the second-best SPLMVC and COPER by 4.3% and 4.2%, respectively. These results demonstrate the effectiveness of combining uncertainty-driven information bottleneck learning, evidential fusion, and directed semantic-evidential alignment for multi-modal clustering.

Parameter Sensitivity Analysis. To assess the parameter sensitivity of the proposed UNIT-MIB, we employed a grid search approach, varying the trade-off hyperparameters α and λ within the range of $[0.1, 1.0]$ at intervals of 0.1. The experimental results across the five benchmark datasets are displayed in Figure 2. As observed, our method performs consistently well across all datasets, with no significant deterioration in clustering performance (ACC) under most parameter combinations. This indicates that our framework is not highly sensitive to hyperparameter variations and demonstrates remarkably stable and robust performance.

Convergence Analysis. To vividly show the convergence of loss terms and clustering performance (ACC and NMI), we conducted experiments across all datasets, with the results presented in Figure 3. From left to right, the five plots correspond to the experimental results on Caltech-2V, Scene15, RGBD, Flickr, and IAPR, respectively. As observed from the curves, the loss values across all five datasets decrease rapidly in the early stages of training, particularly within the first 25 epochs, which indicates that the model efficiently optimizes the objective function and learns the latent semantic structure. Correspondingly, the ACC and NMI values rise steadily as the training progresses and eventually reach a stable plateau after approximately 50 epochs. Specifically, the loss

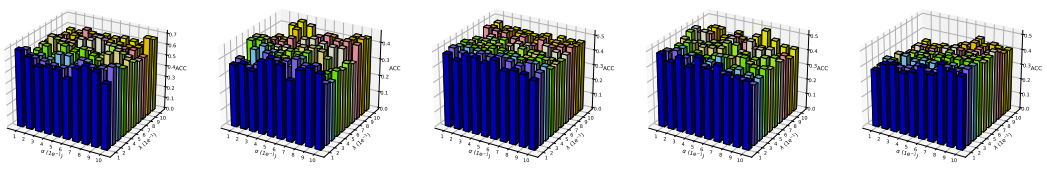

Figure 2: Parameter sensitivity analysis on Caltech-2V, Scene15, RGBD, Flickr, and IAPR.

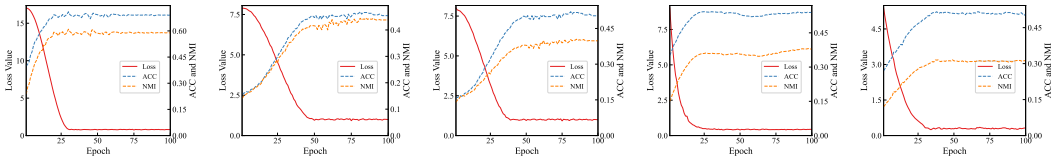

Figure 3: Training process of UNIT-MIB on multiple datasets.

Table 2: The ablation results of the proposed model.

Methods	Caltech-2V		Scene15		RGBD		Flickr		IAPR	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
(1) $\mathcal{L}_{IB}$	53.1	46.1	28.3	21.4	39.4	30.4	43.1	31.6	37.1	19.2
(2) $\mathcal{L}_{IB} + \mathcal{L}_{EDL}$	60.4	47.7	<u>32.8</u>	<u>35.2</u>	42.6	33.1	<u>46.6</u>	<u>32.5</u>	41.3	22.5
(3) $\mathcal{L}_{IB} + \mathcal{L}_{D}$	<u>63.7</u>	<u>51.2</u>	30.3	32.6	<u>46.1</u>	<u>35.9</u>	43.1	27.7	<u>45.1</u>	<u>23.3</u>
(4) UNIT-MIB	**69.3**	**58.4**	**44.3**	**43.6**	**50.2**	**39.8**	**55.9**	**40.2**	**51.9**	**29.7**

curves stabilize at a low value, while the clustering metrics achieve their peak performance early and remain consistent throughout the remaining epochs. These results indicate that UNIT-MIB exhibits stable convergence behavior across the five benchmark datasets.

3.3 Ablation Study

To evaluate the individual contributions of each core component in UNIT-MIB, we conducted extensive ablation experiments across five datasets, with the results summarized in Table 2. Starting from Variant 1, which relies solely on the information bottleneck $\mathcal{L}_{IB}$, we observe that while it captures essential cross-modal structures, its performance remains constrained, yielding only 53.1% ACC on the Caltech-2V dataset.

Integrating the evidential learning module $\mathcal{L}_{EDL}$ in Variant 2 brings substantial improvements, increasing the ACC on Flickr from 43.1% to 46.6% and validating its ability to dynamically suppress noisy modalities and enhance consensus robustness. Similarly, the directed alignment path $\mathcal{L}_{D}$ in Variant 3 significantly improves clustering separability on highly ambiguous datasets like IAPR by achieving a 45.1% ACC, which demonstrates the critical role of Deng mutual information in bridging the semantic-logical gap. Finally, the full UNIT-MIB model achieves peak performance across all metrics by synergistically integrating all components. Notably, on the Flickr dataset, the full model reaches an ACC of 55.9%, substantially outperforming the best reduced variant of 46.6%. On IAPR, the ACC is further improved from 45.1% to 51.9%. This effectively confirms our core motivation: $\mathcal{L}_{EDL}$ provides trustworthy evidence, while $\mathcal{L}_{D}$ guides the representation learning toward a unified and highly discriminative consensus.

4 Conclusion

The network architecture presented in this paper employs a trustworthy multi-modal clustering approach, strategically enhancing performance by integrating uncertainty-driven information bottleneck and evidential dual-path fusion. By adaptively regulating compression intensity and fusion weights, the method effectively mitigates the impact of unreliable modalities. A key component of UNIT-MIB is the directed alignment mechanism based on Deng mutual information, which reduces the discrepancy between semantic representations and evidential consensus. Furthermore, the architecture is highly adaptable, allowing it to be applied to a wide range of clustering tasks with minor modifications, thereby underscoring its practicality and versatility across various real-world scenarios.

In the future, we will focus on optimizing the network architecture to enhance efficiency and reduce computational overhead, while simultaneously exploring the implementation of the proposed method in incomplete data settings. These improvements are expected to significantly enhance the robustness and scalability of the framework in practical applications such as cross-modal retrieval.

Ethics Statement

This paper does not involve any potential ethics issues.

Reproducibility Statement

We have exerted substantial efforts to guarantee the reproducibility of our research. The main manuscript offers explicit descriptions of the proposed framework, algorithmic modules, and evaluation methodologies. Complete proofs of theoretical findings and detailed derivations of key propositions are included in the appendix. For the experimental section, we utilized public datasets and provided thorough accounts of the experimental setups. To further facilitate reproducibility, the full source code and implementation specifics will be made publicly available upon the paper's acceptance.

Statement on the Use of Large Language Models

In accordance with academic integrity guidelines, we hereby declare that no Large Language Models (LLMs) were used in the conceptualization, experimentation, analysis, or writing of this subsection and the related parts of this paper. All content was produced by the authors independently.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Anna Bosch, Andrew Zisserman, and Xavier Muñoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pp. 401–408, 2007.

Matthew Chalk, Olivier Marre, and Gasper Tkacik. Relevant sparse codes with variational information bottleneck. In *Annual Conference on Neural Information Processing Systems*, pp. 1957–1965, 2016.

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

Yong Deng. Deng entropy. *Chaos, Solitons & Fractals*, 91:549–553, 2016.

Zhibin Dong, Dayu Hu, Jiaqi Jin, Siwei Wang, Xinwang Liu, and En Zhu. Selective cross-view topology for deep incomplete multi-view clustering. *IEEE Trans. Image Process.*, 34:4792–4805, 2025.

Ran Eisenberg, Jonathan Svirsky, and Ofir Lindenbaum. COPER: correlation-based permutations for multi-view clustering. In *International Conference on Learning Representations*, 2025.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 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*, pp. 178, 2004.

Yu Geng, Zongbo Han, Changqing Zhang, and Qinghua Hu. Uncertainty-aware multi-view representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7545–7553, 2021.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop on OntoImage*, volume 2, 2006.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(2):2551–2566, 2023.

Shuqing He, Jicong Fan, and Tianwei Yu. G3DC: A gene-graph-guided selective deep clustering method for single cell rna-seq data. *Big Data Min. Anal.*, 7(3):809–827, 2024.

Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, and Yangdong Ye. A survey on information bottleneck. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5325–5344, 2024.

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

Jianqiang Huang, Kai Li, Jinfang Jia, and Xiaoying Wang. Single image super-resolution through image pixel information clustering and generative adversarial network. *Big Data Min. Anal.*, 8 (5):1044–1059, 2025.

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*, pp. 39–43, 2008.

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

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178, 2006.

Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2036–2043, 2009.

Zhengzheng Lou, Chaoyang Zhang, Hang Xue, Yangdong Ye, Qinglei Zhou, and Shizhe Hu. Self-supervised weighted information bottleneck for multi-view clustering. In *International Joint Conference on Artificial Intelligence*, pp. 4643–4650, 2024.

Zhengzheng Lou, Hang Xue, Chaoyang Zhang, and Shizhe Hu. A peer-review look on multi-modal clustering: An information bottleneck realization method. In *International Conference on Machine Learning*, 2025.

Yixuan Luo, Wenming Ma, Xiaolin Du, Qian Xu, Jinghui Zhang, Zhendong Cui, Xiaoyan Dai, and Hongwei Jiang. Deep multi-view clustering via structure preserving learning. *Knowl. Based Syst.*, 346:116096, 2026.

T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.*, 42(3):145–175, 2001.

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.

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.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles — A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, 2002.

Yuan Sun, Yongxiang Li, Zhenwen Ren, Guiduo Duan, Dezhong Peng, and Peng Hu. ROLL: robust noisy pseudo-label learning for multi-view clustering with noisy correspondence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30732–30741, 2025.

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

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*, pp. 1255–1265, 2021.

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*, pp. 16030–16039, 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*, pp. 22957–22966, 2024.

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*, pp. 337–346. ACM, 2023.

Qian Zhang, Lin Zhang, Ran Song, Runmin Cong, Yonghuai Liu, and Wei Zhang. Learning common semantics via optimal transport for contrastive multi-view clustering. *IEEE Trans. Image Process.*, 33:4501–4515, 2024.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International Conference on Machine Learning*, pp. 41753–41769, 2023.

A APPENDIX

In the supplementary material, we provide the following additional details:

- Appendix A.1: Derivation and Analysis of Deng Mutual Information.
- Appendix A.2: Proof of the Variational Upper Bound for Information Compression.
- Appendix A.3: Proof of the Contrastive Alignment Lower Bound.
- Appendix A.4: Dataset Details.
- Appendix A.5: Computational Complexity.

A.1 DERIVATION AND ANALYSIS OF DENG MUTUAL INFORMATION

This section provides additional theoretical details for the Deng mutual information employed in Section 2.6. We first derive the Deng entropy used for subjective opinions, then present the construction of the proposed Deng-entropy-based information measure, and finally explain how directionality is introduced through stop-gradient optimization.

Derivation of Deng Mutual Information. Consider a frame of discernment $\Theta = \{\omega_1, \ldots, \omega_C\}$ containing C mutually exclusive clustering categories. For a subjective opinion $\mathcal{M} = \{b_1, \ldots, b_C, u\}$, the belief masses satisfy

$$\sum_{k=1}^{C} b_k + u = 1,$$

where b_k represents the belief assigned to the singleton category $\{\omega_k\}$ and u represents the uncommitted uncertainty mass associated with the entire frame Θ.

The general Deng entropy of a basic probability assignment $m(A)$ is defined over the non-empty subsets of Θ as

$$H_D(m) = -\sum_{A \in 2^\Theta \setminus \{\emptyset\}} m(A) \log_2 \frac{m(A)}{2^{|A|} - 1}, \tag{13}$$

where terms with $m(A) = 0$ are taken to contribute zero. Excluding the empty set avoids the undefined denominator $2^{|\emptyset|} - 1 = 0$.

In our subjective-opinion formulation, the focal elements consist of the singleton sets $\{\omega_k\}$ and the global uncertainty set Θ. Accordingly, we have $m(\{\omega_k\}) = b_k$ and $m(\Theta) = u$. Since $|\{\omega_k\}| = 1$ and $|\Theta| = C$, Eq. (13) becomes

$$H_D(\mathcal{M}) = -\sum_{k=1}^{C} b_k \log_2 b_k - u \log_2 \left(\frac{u}{2^C - 1} \right). \tag{14}$$

Equivalently, the entropy can be decomposed as

$$H_D(\mathcal{M}) = -\sum_{k=1}^{C} b_k \log_2 b_k - u \log_2 u + u \log_2 (2^C - 1). \tag{15}$$

The first two terms characterize the dispersion of the committed belief and uncertainty masses, while the last term, $u \log_2(2^C - 1)$, explicitly accounts for the cardinality of the global uncertainty focal set Θ. Thus, the uncertainty mass contributes to the information measure through an additional cardinality-dependent term that is absent from Shannon entropy over class probabilities alone. Importantly, this decomposition does not imply that the complete Deng entropy is strictly monotonic with respect to u, since the belief masses and uncertainty mass are jointly constrained by normalization. Rather, it shows that the vacuity-related information carried by u is explicitly represented.

Let

$$\mathcal{M}^S = \{b_1^S, \ldots, b_C^S, u^S\}$$

denote the feature-fused opinion generated from the semantic representation, and let

$$\mathcal{M}^{DS} = \{b_1^{DS}, \ldots, b_C^{DS}, u^{DS}\}$$

denote the evidence-fused consensus obtained through Dempster-Shafer reasoning.

To characterize the discrepancy between these two opinions, we construct their equally weighted component-wise mixture

$$\bar{\mathcal{M}} = \{\bar{b}_1, \ldots, \bar{b}_C, \bar{u}\}, \tag{16}$$

where

$$\bar{b}_k = \frac{b_k^S + b_k^{DS}}{2}, \qquad \bar{u} = \frac{u^S + u^{DS}}{2}.$$

Equivalently, we use the shorthand $\bar{\mathcal{M}} = (\mathcal{M}^S + \mathcal{M}^{DS})/2$. Since both original opinions satisfy the normalization constraint, the mixed opinion also satisfies

$$\sum_{k=1}^{C} \bar{b}_k + \bar{u} = 1.$$

To quantify the discrepancy between the feature-derived opinion $\mathcal{M}^S$ and the evidence-fused consensus $\mathcal{M}^{DS}$, we define the following Deng-entropy-based information measure:

$$I_D(\mathcal{M}^S; \mathcal{M}^{DS}) = H_D(\bar{\mathcal{M}}) - \frac{1}{2} H_D(\mathcal{M}^S) - \frac{1}{2} H_D(\mathcal{M}^{DS}). \tag{17}$$

In this work, we refer to this quantity as Deng mutual information.

Substituting the component-wise mixture into Eq. (17) yields

$$I_D(\mathcal{M}^S; \mathcal{M}^{DS}) = H_D\left(\frac{\mathcal{M}^S + \mathcal{M}^{DS}}{2}\right) - \frac{H_D(\mathcal{M}^S) + H_D(\mathcal{M}^{DS})}{2}. \tag{18}$$

By further substituting the subjective-opinion form of Deng entropy in Eq. (14), the proposed measure can be explicitly written as

$$\begin{aligned} I_D = & -\sum_{k=1}^{C} \bar{b}_k \log_2 \bar{b}_k - \bar{u} \log_2 \left(\frac{\bar{u}}{2^C - 1}\right) \\ & + \frac{1}{2} \sum_{k=1}^{C} b_k^S \log_2 b_k^S + \frac{1}{2} u^S \log_2 \left(\frac{u^S}{2^C - 1}\right) \\ & + \frac{1}{2} \sum_{k=1}^{C} b_k^{DS} \log_2 b_k^{DS} + \frac{1}{2} u^{DS} \log_2 \left(\frac{u^{DS}}{2^C - 1}\right). \end{aligned} \tag{19}$$

Unlike a discrepancy computed solely from class probabilities, this formulation operates directly on subjective opinions and therefore incorporates both committed belief masses and the explicit epistemic uncertainty mass. This property allows the discrepancy between the semantic and evidential opinions to reflect not only class-level disagreement but also differences in their uncertainty states.

Although $I_D(\mathcal{M}^S; \mathcal{M}^{DS})$ is symmetric with respect to its two arguments, the proposed alignment becomes directional during optimization by stopping gradients through the evidential consensus:

$$\mathcal{L}_D = I_D\left(\mathcal{M}^S; \mathrm{sg}(\mathcal{M}^{DS})\right). \tag{20}$$

Here, $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. Therefore, with respect to $\mathcal{L}_D$, $\mathcal{M}^{DS}$ serves as a fixed guidance target, while gradients are propagated only through the semantic branch. The directionality thus arises from asymmetric gradient flow rather than from the mathematical definition of I_D itself.

A.2 Proof of the Variational Upper Bound for Information Compression

We minimize $I(X^m; H^m)$ to filter task-irrelevant redundancy. Direct computation is intractable because the marginal distribution of the latent representation is generally unknown. We therefore adopt the Variational Information Bottleneck (VIB) formulation. For modality m, the encoder maps the input to a mean $\boldsymbol{\mu}^m(x^m) \in \mathbb{R}^D$ and a diagonal log-variance $\log \boldsymbol{\sigma}^{m,2}(x^m) \in \mathbb{R}^D$. The latent representation is sampled using the reparameterization trick:

$$h^m = \boldsymbol{\mu}^m(x^m) + \boldsymbol{\sigma}^m(x^m) \odot \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{21}$$

which induces the stochastic encoder $q_\phi(h^m|x^m) = \mathcal{N}\big(\boldsymbol{\mu}^m(x^m), \mathrm{diag}(\boldsymbol{\sigma}^{m,2}(x^m))\big)$.

The mutual information between X^m and H^m can be expressed as

$$I(X^m; H^m) = \mathbb{E}_{p(x^m)} \left[\mathrm{KL} \left(q_\phi(h^m|x^m) \,\|\, q_\phi(h^m) \right) \right], \tag{22}$$

where the aggregated posterior is

$$q_\phi(h^m) = \int q_\phi(h^m|x^m) p(x^m)\, dx^m. \tag{23}$$

Since $q_\phi(h^m)$ is generally intractable, we introduce a fixed variational prior $r(h^m) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The expected KL divergence to this prior can be decomposed as

$$\begin{aligned} &\mathbb{E}_{p(x^m)} \left[\mathrm{KL} \left(q_\phi(h^m|x^m) \,\|\, r(h^m) \right) \right] \\ &\quad = I(X^m; H^m) + \mathrm{KL} \left(q_\phi(h^m) \,\|\, r(h^m) \right). \end{aligned} \tag{24}$$

Therefore,

$$\begin{aligned} I(X^m; H^m) = \mathbb{E}_{p(x^m)} &\left[\mathrm{KL} \left(q_\phi(h^m|x^m) \,\|\, r(h^m) \right) \right] \\ &- \mathrm{KL} \left(q_\phi(h^m) \,\|\, r(h^m) \right). \end{aligned} \tag{25}$$

By the non-negativity of the KL divergence,

$$I(X^m; H^m) \le \mathbb{E}_{p(x^m)} \left[\mathrm{KL} \left(q_\phi(h^m|x^m) \,\|\, r(h^m) \right) \right]. \tag{26}$$

Since both $q_\phi(h^m|x^m)$ and $r(h^m)$ are Gaussian, the KL divergence admits the closed-form expression

$$\begin{aligned} &\mathrm{KL} \left(\mathcal{N}(\boldsymbol{\mu}^m, \mathrm{diag}(\boldsymbol{\sigma}^{m,2})) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I}) \right) \\ &\quad = -\frac{1}{2} \sum_{d=1}^{D} \left(1 + \log \sigma_d^{m,2} - (\mu_d^m)^2 - \sigma_d^{m,2} \right). \end{aligned} \tag{27}$$

Approximating the expectation over $p(x^m)$ using a mini-batch of B samples yields the tractable VIB compression objective

$$\mathcal{L}^m_{\mathrm{comp}} = \frac{1}{B} \sum_{i=1}^{B} \mathrm{KL} \left(q_\phi(h_i^m|x_i^m) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I}) \right). \tag{28}$$

Thus, minimizing $\mathcal{L}^m_{\mathrm{comp}}$ minimizes a variational upper bound on $I(X^m; H^m)$, providing a tractable realization of the information-compression component in UNIT-MIB.

A.3 Proof of the Contrastive Alignment Lower Bound

The cross-modal alignment term in UNIT-MIB aims to preserve the shared semantic information between modality-specific representations by maximizing the mutual information $I(H^u; H^v)$. However, directly evaluating this mutual information is generally intractable because the joint and marginal distributions of the learned representations are unknown. We therefore employ the InfoNCE principle to obtain a tractable variational lower bound and show that the NT-Xent objective used for cross-modal alignment maximizes this bound.

Contrastive Sampling. Consider an anchor representation H^u from modality u and B candidate representations $\{H_1^v, \dots, H_B^v\}$ from modality v, where B denotes the mini-batch size. Let $I \in \{1, \dots, B\}$ denote the index of the positive sample and assume that I is uniformly distributed:

$$p(I = i) = \frac{1}{B}.$$

Conditioned on an anchor h^u and a positive index $I = i$, the positive representation h_i^v is sampled from the conditional distribution $p(h^v|h^u)$, while the remaining $B-1$ candidates are independently sampled from the marginal distribution $p(h^v)$. Therefore, the joint distribution of the contrastive sampling process can be written as

$$p(i, h^u, h_{1:B}^v) = \frac{1}{B} p(h^u) p(h_i^v|h^u) \prod_{j \neq i} p(h_j^v). \tag{29}$$

We further define a reference distribution in which all B candidate representations are independently sampled from the marginal distribution:

$$q(i, h^u, h_{1:B}^v) = \frac{1}{B} p(h^u) \prod_{j=1}^{B} p(h_j^v). \tag{30}$$

Taking the ratio between Eq. (29) and Eq. (30) gives

$$\frac{p(i, h^u, h_{1:B}^v)}{q(i, h^u, h_{1:B}^v)} = \frac{p(h_i^v|h^u)}{p(h_i^v)}. \tag{31}$$

Hence, the KL divergence between these two distributions becomes

$$\begin{aligned} &D_{KL}\left(p(i, h^u, h_{1:B}^v) \parallel q(i, h^u, h_{1:B}^v)\right) \\ &= \mathbb{E}\left[\log \frac{p(h_I^v|h^u)}{p(h_I^v)}\right] = I(H^u; H^v). \end{aligned} \tag{32}$$

Therefore, the mutual information between the two modality-specific representations can be interpreted as the KL divergence between the positive-sampling distribution and the corresponding independent reference distribution.

Positive-index Classification. Let $\mathcal{Z} = (H^u, H_{1:B}^v)$ denote the complete set of observations used in the contrastive task. Using Bayes' rule together with Eq. (29), the posterior probability that the i-th candidate is the positive sample is

$$p(I = i|\mathcal{Z}) = \frac{\frac{p(h_i^v|h^u)}{p(h_i^v)}}{\sum_{j=1}^{B} \frac{p(h_j^v|h^u)}{p(h_j^v)}}. \tag{33}$$

Equation (33) shows that the Bayes-optimal contrastive discriminator depends on the density ratio

$$r(h^u, h^v) = \frac{p(h^v|h^u)}{p(h^v)}. \tag{34}$$

Since this density ratio is unknown in practice, we approximate its logarithm using a critic function. In UNIT-MIB, the critic is instantiated by the temperature-scaled representation similarity:

$$f_\theta(h^u, h^v) = \frac{\mathrm{sim}(h^u, h^v)}{\tau}, \tag{35}$$

where τ denotes the temperature parameter. The corresponding model posterior over the positive index is defined as

$$q_\theta(I = i|\mathcal{Z}) = \frac{\exp(f_\theta(h^u, h_i^v))}{\sum_{j=1}^{B} \exp(f_\theta(h^u, h_j^v))}. \tag{36}$$

Thus, contrastive learning can be viewed as a B-way classification problem in which the model attempts to distinguish the joint sample from candidates drawn from the product of marginals.

Mutual-information Lower Bound. We now establish the relationship between this classification objective and $I(H^u; H^v)$. From Eq. (32), we have

$$I(H^u; H^v) = D_{KL}\left(p(I, \mathcal{Z}) \parallel p(I)q(\mathcal{Z})\right), \tag{37}$$

where

$$q(\mathcal{Z}) = p(h^u) \prod_{j=1}^{B} p(h_j^v)$$

denotes the independent reference distribution.

Using the chain rule of KL divergence, this quantity can be decomposed as

$$I(H^u; H^v) = I(I; \mathcal{Z}) + D_{KL}\left(p(\mathcal{Z}) \parallel q(\mathcal{Z})\right). \tag{38}$$

Since KL divergence is non-negative,

$$I(H^u; H^v) \geq I(I; \mathcal{Z}). \tag{39}$$

Because the positive index is uniformly distributed, its entropy is

$$H(I) = \log B.$$

Therefore,

$$I(I; \mathcal{Z}) = \log B - H(I|\mathcal{Z}). \tag{40}$$

The contrastive objective minimizes the cross-entropy between the true posterior $p(I|\mathcal{Z})$ and the model posterior $q_\theta(I|\mathcal{Z})$:

$$\mathcal{L}_{\text{NCE}} = \mathbb{E}_{p(I,\mathcal{Z})}\left[-\log q_\theta(I|\mathcal{Z})\right]. \tag{41}$$

The cross-entropy can be decomposed into the conditional entropy and a KL divergence:

$$\mathcal{L}_{\text{NCE}} = H(I|\mathcal{Z}) + \mathbb{E}_{p(\mathcal{Z})}\left[D_{KL}\left(p(I|\mathcal{Z}) \parallel q_\theta(I|\mathcal{Z})\right)\right]. \tag{42}$$

By the non-negativity of KL divergence,

$$H(I|\mathcal{Z}) \leq \mathcal{L}_{\text{NCE}}. \tag{43}$$

Substituting Eq. (43) into Eq. (40) gives

$$I(I; \mathcal{Z}) \geq \log B - \mathcal{L}_{\text{NCE}}. \tag{44}$$

Combining Eq. (39) and Eq. (44), we obtain the InfoNCE lower bound:

$$I(H^u; H^v) \geq \log B - \mathcal{L}_{\text{NCE}}. \tag{45}$$

Connection to the NT-Xent Objective. With the critic defined in Eq. (35), the empirical contrastive loss for a mini-batch $\{(h_i^u, h_i^v)\}_{i=1}^B$ becomes

$$\mathcal{L}_{\text{NT-Xent}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp\left(\text{sim}(h_i^u, h_i^v)/\tau\right)}{\sum_{j=1}^{B} \exp\left(\text{sim}(h_i^u, h_j^v)/\tau\right)}. \tag{46}$$

This objective is an empirical realization of $\mathcal{L}_{\text{NCE}}$. Consequently,

$$I(H^u; H^v) \geq \log B - \mathcal{L}_{\text{NT-Xent}}. \tag{47}$$

For a fixed mini-batch size B, $\log B$ is constant. Therefore, minimizing $\mathcal{L}_{\text{NT-Xent}}$ maximizes a variational lower bound on the cross-modal mutual information $I(H^u; H^v)$. In combination with the VIB compression objective, this establishes the information-theoretic role of the cross-modal alignment component in UNIT-MIB: VIB suppresses modality-specific redundant information, while NT-Xent encourages different modality-specific representations to retain their shared semantic information.

Extension to Evidential Guidance. The same InfoNCE lower-bound formulation is further employed to instantiate the evidential guidance term $I(H^m; \mathcal{M}^{DS})$ in $\mathcal{L}_{IB}$. Since the complete DS consensus is represented by the subjective opinion $\mathcal{M}^{DS} = \{b_1^{DS}, \ldots, b_C^{DS}, u^{DS}\}$, we use its belief-mass component $b^{DS} \in \mathbb{R}^C$ as the contrastive representation of the evidential consensus. Consistent with the implementation, the belief vector is detached from the evidential branch and then mapped into the D-dimensional latent space through a learnable projection head:

$$\tilde{z}^{DS} = g_\psi\big(\mathrm{sg}(b^{DS})\big) \in \mathbb{R}^D, \tag{48}$$

where $g_\psi(\cdot)$ denotes the learnable projection function and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator.

The projected evidential representation $\tilde{z}^{DS}$ is then contrasted with each modality-specific representation H^m. Following the same derivation above, the corresponding mutual information admits the lower bound

$$I(H^m; \tilde{z}^{DS}) \geq \log B - \mathcal{L}_{\text{NT-Xent}}(H^m, \tilde{z}^{DS}). \tag{49}$$

Therefore, minimizing the corresponding NT-Xent objective maximizes a tractable lower bound on the mutual information between each modality-specific representation and the projected evidential consensus. The stop-gradient operation prevents the contrastive guidance objective from directly modifying the DS belief representation, while the learnable projection head g_ψ and the modality-specific representation branch remain optimizable. In this manner, the evidential consensus provides stable guidance for representation learning without being reversely affected by the evidential-guidance objective.

A.4 Dataset Details

We evaluate UNIT-MIB on five public multi-modal datasets. Caltech-2V contains 1,400 images from 7 classes and is represented by Wavelet moments and CENTRIST features. RGBD consists of 1,449 samples from 13 indoor scene categories, represented by RGB images and depth information. Scene15 contains 4,485 images from 15 scene categories and is represented by GIST, PHOG, and LBP features. For IAPR, we use a filtered subset of 7,855 images from 6 categories, where each sample has at least four labels and is represented by visual features and textual descriptions. Flickr contains 12,154 denoised images from 6 categories and is represented by visual and textual modalities.

Table 3: Statistics of the datasets used in the experiments.

Dataset	Modalities	Samples	Clusters
Caltech-2V	2	1,400	7
RGBD	2	1,449	13
Scene15	3	4,485	15
IAPR	2	7,855	6
Flickr	2	12,154	6

A.5 Computational Complexity

We analyze the computational complexity of UNIT-MIB according to its main components. Let N denote the total number of samples, B the mini-batch size, M the number of modalities, C the number of clusters, and D the dimensionality of the latent representation. Since the specific encoder architecture may vary across modalities, we denote the computational cost of all modality-specific encoders for one mini-batch by T_{enc} and focus on the additional complexity introduced by the proposed framework.

Uncertainty-driven information bottleneck. For the VIB compression term $I(X^m; H^m)$, the KL divergence is evaluated element-wise over the D-dimensional latent representation for each sample and modality, resulting in a complexity of

$$O(BMD).$$

For the cross-modal information-preservation term, UNIT-MIB evaluates contrastive alignment between every pair of modalities. There are $M(M-1)/2$ modality pairs, and constructing an $B \times B$

similarity matrix for each pair requires $O(B^2D)$. Therefore, the pairwise cross-modal alignment has complexity

$$O(M^2B^2D).$$

The uncertainty-weighted evidential guidance between each modality representation H^m and the global consensus $\mathcal{M}^{DS}$ introduces an additional contrastive alignment for M modalities, yielding

$$O(MB^2D).$$

Thus, the information-bottleneck component has the overall complexity

$$T_{IB} = T_{\text{enc}} + O(BMD) + O(M^2B^2D) + O(MB^2D).$$

Uncertainty-weighted feature fusion. The semantic fusion path constructs the global representation $H^S = \sum_{m=1}^{M} w(u^m)H^m$. Computing the uncertainty-aware weights and aggregating M modality-specific representations require

$$O(BMD),$$

which is linear in the number of modalities and the mini-batch size.

Evidential learning and DS fusion. UNIT-MIB constructs subjective opinions for the M modality-specific representations and the fused representation H^S. Each evidential classifier maps a D-dimensional representation to C evidence values. Therefore, the $M + 1$ evidential branches require

$$O\big(B(M+1)DC\big).$$

The computation of belief masses, epistemic uncertainties, and the evidential learning objective is linear in the number of clusters, resulting in

$$O\big(B(M+1)C\big).$$

For DS fusion, the conflict coefficient between two opinions is computed over pairs of distinct cluster beliefs. According to the combination rule used in UNIT-MIB, a direct computation requires $O(C^2)$ operations for each sample. Since the M modality-specific opinions are recursively combined, the overall DS fusion complexity is

$$O(BMC^2).$$

Directed alignment via Deng mutual information. The directed alignment operates on the feature-fused opinion $\mathcal{M}^S$ and the evidence-fused opinion $\mathcal{M}^{DS}$, each consisting of C belief masses and one uncertainty mass. Computing Deng entropy and the corresponding Deng mutual information therefore requires

$$O(BC).$$

The stop-gradient operation applied to $\mathcal{M}^{DS}$ does not introduce additional asymptotic computational complexity.

Combining the above components, the computational complexity for one mini-batch is

$$T_{\text{batch}} = T_{\text{enc}} + O(M^2B^2D) + O(MB^2D) + O\big(B(M+1)DC\big) + O(BMC^2).$$

For E training epochs, with approximately N/B mini-batches per epoch, the total training complexity is

$$T_{\text{total}} = O\left(E\frac{N}{B}T_{\text{batch}}\right).$$

In practice, the dominant computation arises from the pairwise instance-level contrastive alignment, whose complexity scales quadratically with the mini-batch size as $O(M^2B^2D)$. In contrast, the distinctive trustworthy components of UNIT-MIB, including uncertainty-weighted feature fusion, evidential opinion modeling, DS fusion, and Deng mutual-information alignment, scale linearly with the mini-batch size. Therefore, these modules introduce moderate additional overhead compared with the contrastive alignment operations.