Calibrating Multimodal Learning

Huan Ma * [1 2] **Qingyang Zhang** * [1] **Changqing Zhang** [1 3]
Bingzhe Wu [2] **Huazhu Fu** [4] **Joey Tianyi Zhou** [4 5] **Qinghua Hu** [1 3]

Abstract

Multimodal machine learning has achieved remarkable progress in a wide range of scenarios. However, the reliability of multimodal learning remains largely unexplored. In this paper, through extensive empirical studies, we identify current multimodal classification methods suffer from unreliable predictive confidence that tend to rely on partial modalities when estimating confidence. Specifically, we find that the confidence estimated by current models could even increase when some modalities are corrupted. To address the issue, we introduce an intuitive principle for multimodal learning, i.e., the confidence should not increase when one modality is removed. Accordingly, we propose a novel regularization technique, i.e., Calibrating Multimodal Learning (CML) regularization, to calibrate the predictive confidence of previous methods. This technique could be flexibly equipped by existing models and improve the performance in terms of confidence calibration, classification accuracy, and model robustness.

1. Introduction

Multimodal data widely exist in real-world applications such as medical analysis (Perrin et al., 2009), social media (Wang et al., 2019), and autonomous driving (Khodayari et al., 2010). To fully explore the potential value of each modality, multimodal learning emerges as a promising way to train a machine learning (ML) model by integrating all available multimodal cues for further data analysis tasks. Numerous approaches have been proposed to build multimodal learning paradigms for various tasks (Wang et al., 2019; Antol et al., 2015; Bagher Zadeh et al., 2018; Kishi et al., 2019). Despite above progresses, the reliability of current multimodal learning methods remains largely unexplored. In the setting of classification, one key aspect of the reliability is to build a high-quality confidence estimator (Moon et al., 2020; Corbière et al., 2019; Guo et al., 2017), which can quantitatively characterize the probability that predictions will be correct. With such an estimator, further processing can be taken to improve the performance of the system (e.g., human assistance) when the predictive uncertainty is high. This is especially useful in high-stake scenarios (Hafner et al., 2019; Qaddoum & Hines, 2012).

In the setting of multimodal learning, in addition to exact overall prediction confidence, the relationship between the confidence and the number of modalities should also be taken into concerns. Intuitively, the confidence of an ideal multimodal classifier should not increase when one modality is removed (for brevity, we initialize the question as "one modality", and the same phenomenon is observed when removing more than one modality). An illustrative example of an ideal confidence estimator is shown in Fig. 1, where the confidence gradually decreases when the observed information becomes less comprehensive. However, we conduct extensive empirical studies on current methods and observe that when one modality is removed, the overall confidence estimated by them can even increase. This observation contradicts the common assumption of multimodal learning since modalities are assumed to be predictive of the target for most multimodal learning tasks (Wu et al., 2022) and the principle "*the essence of information is to eliminate uncertainty (Shannon)*" in informatics (Soni & Goodman, 2017; Burgin, 2002). Intuitively, this implies that the models are more inclined to believe in a unique modality and is prone to be affected by this modality, which has also been shown in prior works (Wu et al., 2022; Wang et al., 2020). This further impairs the robustness of the learned models, i.e., the models are easy to be influenced when some modalities are corrupted, since the models can not make decisions according to a trustworthy confidence (probability) estimator.

A natural idea to address the above issue is to employ re-

[*] Equal contribution [1] College of Intelligence and Computing, Tianjin University, Tianjin, China [2] AI Lab, Tencent, Shenzhen, China [3] Tianjin Key Lab of Machine Learning, Tianjin, China [4] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore [5] Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore. Correspondence to: Changqing Zhang <zhanchangqing@tju.edu.cn>, Bingzhe Wu <bingzhewu@tencent.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

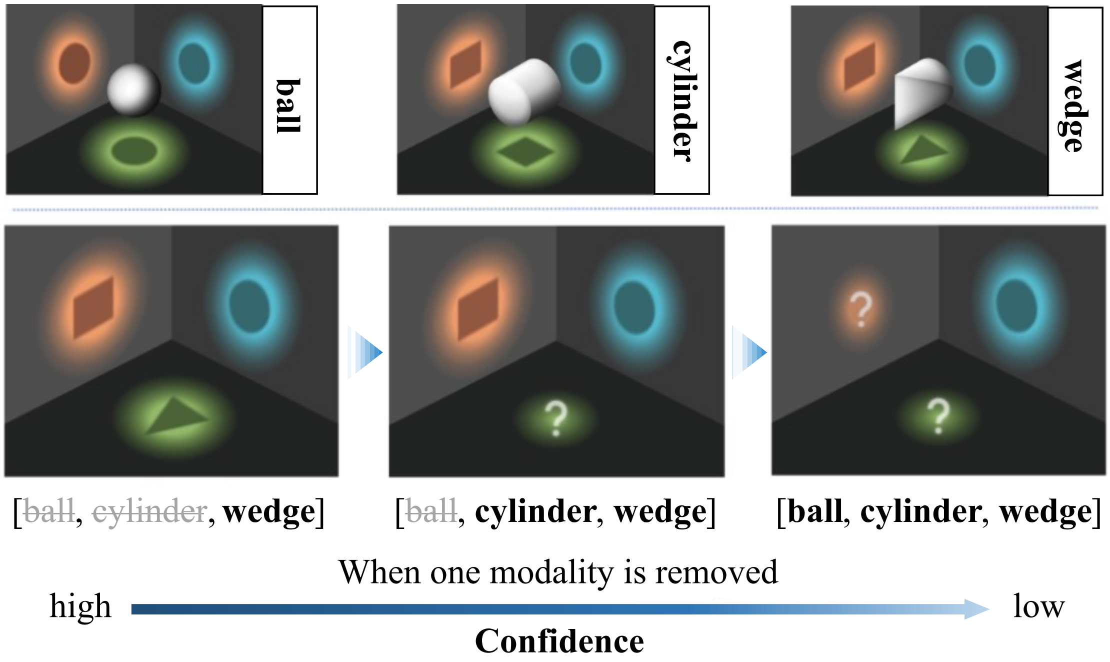

ball
cylinder
wedge
[~~ball~~, ~~cylinder~~, wedge]
[~~ball~~, cylinder, wedge]
[ball, cylinder, wedge]
When one modality is removed
high
low
Confidence

Figure 1: Motivation of calibrating multimodal learning. The confidence of an ideal multimodal classifier should decrease or at least not increase when one modality is removed (even when the removed modality is noised, or it indicates the model takes noise as semantics and the model is not trustworthy).

cent uncertainty calibration methods such as temperature scaling (Guo et al., 2017) or Bayesian learning (Cobb & Jalaian, 2021; Karaletsos & Bui, 2020; Foong et al., 2020), which can build more accurate confidence estimation than the traditional training/inference manner. However, these approaches do not explicitly consider the relationship between different modalities (i.e., they can only calibrate the overall confidence but can not calibrate the confidence of using different number of modalities) and thus still fail to achieve satisfactory performance in the multimodal learning setting. To address this issue, we propose a novel regularization technique called **C**alibrating **M**ultimodal **L**earning (CML) which enforces the consistency between prediction confidence and the number of modalities. The motivation of CML is based on a natural intuition, i.e., the prediction confidence should decrease (at least not increase) when one modality is removed, which could intrinsically improve the confidence calibration. Specifically, we propose a simple regularization term that enforces a model to learn an intuitive ranking relationship by adding a penalty for the samples whose predictive confidence will increase when one modality is removed. The main contributions of this paper are summarized as follows:

- We conduct extensive empirical studies to show that most existing multimodal learning paradigms tend to be over-confident on partial modalities (different samples are over-confident on different modalities rather than all samples are over-confident on the same modalities), which implies that they fail to achieve trustworthy confidence estimation.
- We introduce a measure to evaluate the reliability of the confidence estimation from the confidence ranking perspective, which can characterize whether a multimodal learning method can treat all modalities fairly.
- We propose a regularization strategy to calibrate the confidence of various multimodal learning methods, and then conduct extensive experiments to show the superiority of our method in terms of the confidence calibration (Table 1), classification accuracy (Table 2) and model robustness (Table 3).

2. Related Work

Uncertainty estimation provides a way for trustworthy prediction (Abdar et al., 2021; Chau et al., 2021; Slack et al., 2021; Singh et al., 2021; Ning et al., 2021; Zhang et al., 2021). Uncertainty can be used as an indicator of whether the predictions given by models are prone to be wrong (Ritter et al., 2021; Wang & Zou, 2021; Zaidi et al., 2021; Stadler et al., 2021; Bai et al., 2021; Rahaman & thiery, 2021; Galil & El-Yaniv, 2021; Upadhyay et al., 2021). Many uncertainty-based models have been proposed in the past decades, such as Bayesian neural networks (Neal, 2012; MacKay, 1992; Denker & LeCun, 1990; Kendall & Gal, 2017), Dropout (Molchanov et al., 2017), Deep ensembles (Lakshminarayanan et al., 2017; Havasi et al., 2020), and DUQ (van Amersfoort et al., 2020) built upon RBF networks. **Prediction confidence** (Sahoo et al., 2021; Wald et al., 2021; Pan et al., 2021; Luo et al., 2021; Xu et al., 2021; Chung et al., 2021; Xiong et al., 2021) is always referred to in classification models, which expects the predicted class probability to be consistent with the empirical accuracy (Qin et al., 2021; Minderer et al., 2021; Zhao et al., 2021; Tian et al., 2021; Karandikar et al., 2021; Jeong et al., 2021). Many methods focus on smoothing the prediction probabilities distribution, such as Label smoothing (Müller et al., 2019), focal loss (Mukhoti et al., 2020), TCP (Corbière et al., 2019)and Temperature scaling (TS) (Guo et al., 2017). More related researches please refer to Appendix G.

Multimodal learning emerges as a promising way to exploit complementary information from different modalities. How to benefit from multimodal data has been a popular research direction, and researchers usually focus on improving architectural designs of the multimodal model (Pérez-Rúa et al., 2019; Sun et al., 2021). In the setting of multimodal classification, MMTM (Joze et al., 2020) achieves state-of-the-art performance by connecting corresponding convolutional layers from different uni-modal branches. Considering the proposed method calibrating confidence with using different number of modalities, multimodal classifiers that can deal with incomplete data are natural candidates to validate our motivation. There is a wide range of research interests in handling missing modalities for multimodal learning, including imputation-independent methods (Zhang et al., 2019) and imputation-dependent methods (Mattei & Frellsen, 2019; Wu & Goodman, 2018). For imputation-independent methods, there is no need to reconstruct the missing modalities and conduct classification using the imputed data. Imputation-dependent methods usually conduct classification with two stages, reconstructing the missing modalities and making classification according to the reconstructed modalities. In this paper, we employ CPM-Nets (Zhang et al., 2019), MIWAE (Mattei & Frellsen, 2019), and MMTM (Joze et al., 2020) to validate our motivation due to their representativeness in multimodal learning.

3. Method

In this section, we first introduce some basic notations in Section 3.1. We show the basic assumption of our method and its empirical motivation in Section 3.2 based on the principle "the essence of information is to eliminate uncertainty", and then evaluate the confidence estimation performance of current multimodal methods in Section 3.3 and find they violate the principle. At the end, we propose a simple yet effective regularization technique to improve the confidence estimation of multimodal models and elaborate the technical details in Section 3.4.

3.1. Notation

We define the training data as $\mathcal{D} = \left\{\{x_i^m\}_{m=1}^M, y_i\right\}_{i=1}^N$, where x_i^m is the m-th modality of the i-th sample, and $y_i \in \{1, \cdots, K\}$ is the corresponding class label. To distinguish one modality or a set of modalities, we use x^m and $x^{(\mathbb{S})}$ to represent the m-th modality and multiple modalities respectively, where $\mathbb{S}$ is a set of modalities' indexes (e.g., if we have $\mathbb{S} = \{1, 2\}$, then $x^{(\mathbb{S})}$ indicates a feature set consisting of x^1 and x^2, and $x^{(\mathbb{M})} = \{x^1, \cdots, x^M\}$ indicates the complete M modalities). The goal is to learn a function parameterized by θ: $f(x^{(\mathbb{M})}, \theta) \to z$, where the output z of the network is a vector of K values called logits. Then the logits vector is transformed by a softmax layer: $\hat{p_k} = e^{z_k} / \sum_k e^{z_k}$, where the probability distribution of a sample x is defined as $\mathrm{P}(y \mid \theta, x^{(\mathbb{M})}) = \{\hat{p_k}\}_1^K$. The predicted class label is $\hat{y} = \arg\max_y \mathrm{P}(y \mid \theta, x^{(\mathbb{M})})$, and the confidence is defined as $\mathrm{Conf}(x^{(\mathbb{M})}) = \max_y \mathrm{P}(y \mid \theta, x^{(\mathbb{M})})$.

3.2. Basic Assumption

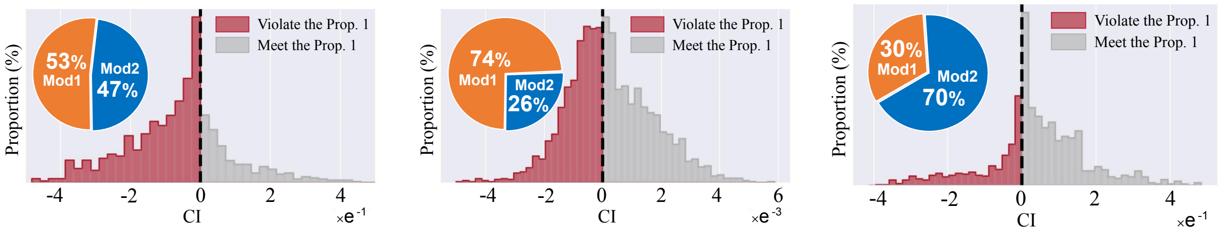
In real-world applications, the quality of multimodal data is usually unstable (e.g., some modalities may be corrupted), so the quality of the multimodal input should be reflected in some quantitative manner (i.e., predictive confidence) which is especially important when multimodal models are deployed for the high-stake tasks. However, it is difficult to exactly define the "quality" of each sample, and we can not define the exact functional relationship between the quality and confidence since the confidence from different models is basically different for a same sample. This issue results in the lack of supervision for confidence estimation. Fortunately, according to the principle "*the essence of information is to eliminate uncertainty (Shannon)*" in informatics (Soni & Goodman, 2017; Burgin, 2002) (i.e., more information, less uncertainty), we can approximate this relationship with a ranking-based form as follow:

> **Proposition 3.1.** *Given two versions of a sample $x^{(\mathbb{M})}$, i.e., $x^{(\mathbb{T})}$ and $x^{(\mathbb{S})}$, if we can assure $\mathbb{T} \subset \mathbb{S} \subseteq \mathbb{M}$, then, for a trustworthy multimodal classifier $f(\cdot)$, it should hold* $\mathrm{Conf}(f(x^{(\mathbb{T})})) \leq \mathrm{Conf}(f(x^{(\mathbb{S})}))$.

For most multimodal learning tasks, all modalities are assumed to be predictive for the target (Wu et al., 2022), and the proposed method is also based on this assumption. For a trustworthy classifier, the predictive confidence should not increase when one modality is removed. We further define the prediction **C**onfidence **I**ncrement (CI) with informativeness increment for a sample as:

$$\mathrm{CI}(x^{(\mathbb{T})}, x^{(\mathbb{S})}) = \mathrm{Conf}(f(x^{(\mathbb{S})})) - \mathrm{Conf}(f(x^{(\mathbb{T})})) \quad \text{s.t. } \mathbb{T} \subset \mathbb{S} \subseteq \mathbb{M}, \tag{1}$$

where $\mathbb{T}$ and $\mathbb{S}$ are sets of modalities' indexes. Specifically, a



(a) MMTM

(b) CPM-Nets

(c) MIWAE

Figure 2: Current methods (Wu et al., 2022; Zhang et al., 2019; Mattei & Frellsen, 2019) violate the Proposition 3.1 (red color indicates the proportion of test samples whose predictive confidence given by the model decreases while providing more modalities, “CI" is defined in Eq. 1). We estimate the performance on two-modality datasets, and the pie charts show that different samples over-rely on different modalities rather than all samples over-rely on the same modality (e.g., “53% Mod1” indicates “among the samples who violate Proposition 3.1, there is 53 percent of samples whose confidence will increase when Mod2 is removed and the other samples will increase confidence when Mod1 is removed”).

negative value indicates a poor confidence estimation performance where the predictive confidence increases when one modality is removed. To quantify the extent that a learned model violates Proposition 3.1, we introduce a novel measure: **V**iolating **R**anking **R**ate (VRR) as the proportion of test samples whose predictive confidence will increase when removing one modality:

$$\mathrm{VRR} = \mathbb{E}_{(\mathbb{T},\,\mathbb{S})}\left[\mathbb{1}\left(\mathrm{CI}(x^{(\mathbb{T})}, x^{(\mathbb{S})}) < 0\right)\right] \quad \text{s.t. } \mathbb{T} \subset \mathbb{S} \subseteq \mathbb{M}. \tag{2}$$

Inspired by prior methods (Moon et al., 2020; Toneva et al., 2018), we initialize $\mathbb{S}$ as the complete modalities, and obtain $\mathbb{T}$ by randomly removing a modality from $\mathbb{S}$. Then $\mathbb{T}$ is regarded as $\mathbb{S}$ for another confidence ranking pair and we repeat this process until there is only one modality remained in $\mathbb{T}$ (Please refer to Appendix A for detail). A natural question then arises: how about the confidence estimation performance of the current methods when one modality is removed?

3.3. Confidence Estimation Performance of Current Multimodal Methods

To evaluate the quality of confidence estimation of existing multimodal classifiers, we compute the VRR score of CPM-Nets (Zhang et al., 2019) and MIWAE (Mattei & Frellsen, 2019), which are two typical methods in handling incomplete multimodal data. In addition to classifiers for incomplete multimodal data, we also evaluate MMTM (Wu et al., 2022), which is a state-of-the-art multimodal classification method. As shown in Table 1, the VRR scores of previous methods are quite high which indicates the prediction confidence on a large portion of samples will violate Proposition 3.1. The visualization is shown in Fig. 2, where the red color indicates the proportion of test samples whose predictive confidence estimated by the model decreases while providing more modalities.

A naive strategy is to re-balance the contribution of every modality (i.e., allocating a smaller weight to the modality that samples are over-confident on during the fusion). As shown in Fig. 2, however, we find that different samples are over-confident on different modalities rather than all samples are over-confident on the same modality. This indicates that the problem can not be solved by re-weighting the overall contribution of different modalities since it will make the confidence estimation of some samples worse. Instead, our method characterizes the relationship between the modalities in sample-wise manner, which inherently calibrates the contribution for all samples. Intuitively, it is risky for a model which usually increases the prediction confidence when one modality is removed, since this usually implies that the confidence of the sample and its informativeness are not matched. For this issue, these models can not be deployed into risk-sensitive applications such as medical diagnosis. As a comparison, our method can significantly decrease VRR score (see more details in Table 1) implying a more trustworthy confidence estimation.

3.4. Calibrating Multimodal Classification Model

As shown in Section 3.3, current multimodal methods usually increase the prediction confidence when one modality is removed, which potentially harms both trustworthiness and performance. To address this issue, the direct strategy is to minimize the following confidence difference:

$$\mathcal{L}^{(\mathbb{T},\,\mathbb{S})} = \mathrm{Conf}(x^{(\mathbb{T})}) - \mathrm{Conf}(x^{(\mathbb{S})}). \tag{3}$$

However, models sometimes can still make an accurate prediction confidently when one modality is removed in practice. Eq. 3 forces models to produce relatively small confidence when one modality is removed, which results in extremely small confidence for each modality (Please refer to Appendix B.6 for detail). For this issue, we relax this regularization by only penalizing the situation that the estimated confidence increases when one modality is removed. For any pair of multimodal inputs which satisfies that $\mathbb{T} \subset \mathbb{S} \subseteq \mathbb{M}$, the regularization can be written as:

$$\mathcal{L}^{(\mathbb{T},\, \mathbb{S})} = \max\left(0, \mathrm{Conf}(x^{(\mathbb{T})}) - \mathrm{Conf}(x^{(\mathbb{S})})\right). \quad (4)$$

For each sample, the total regularization loss is integrated over all pairs of inputs with different numbers of modalities, which is formalized as:

$$\mathcal{L}^{\mathrm{CML}} = \sum_{(\mathbb{T},\, \mathbb{S})} \mathcal{L}^{(\mathbb{T},\, \mathbb{S})}, \quad \{\forall (\mathbb{T},\, \mathbb{S}) | \mathbb{T} \subset \mathbb{S} \subseteq \mathbb{M}\}. \quad (5)$$

The exact computation of above loss needs to enumerate all modality set pairs $(\mathbb{T}, \mathbb{S})$, which is typically computational expensive sometimes. Therefore, we propose to approximate this loss by sampling and it works well in practice. Specifically, we conduct sampling as same as that in computing VRR defined in Eq. 2.

The proposed regularization is general and thus can be equipped by current multimodal classifiers to calibrate their confidence estimation as an additional loss item. We typically provide examples in utilizing the proposed technique in imputation-independent method (i.e., CPM-Nets (Zhang et al., 2019)), imputation-dependent method (i.e., MIWAE (Mattei & Frellsen, 2019)), and recent multimodal classification method (i.e., MMTM (Wu et al., 2022)). The proposed regularization can be deployed to current multimodal methods flexibly, and accordingly the objective function is induced as:

$$\mathcal{L} = \mathcal{L}^{\mathrm{CL}} + \lambda \mathcal{L}^{\mathrm{CML}}, \quad (6)$$

where $\mathcal{L}^{\mathrm{CL}}$ is the classification loss criterion (e.g., cross-entropy loss), and λ is hyperparameter controlling the strength of CML regularization. The process of calibrating multimodal classification are shown in Algorithm 1.

3.5. Discussion and Analyses

○ **Why should a model meet the ranking relationship regardless of class labels?** For multimodal learning, all modalities are assumed to be predictive of the target (Wu et al., 2022), which can be expressed as $I(y, x^m) \geq 0$, where $I(\cdot)$ denotes mutual information (Blum & Mitchell, 1998) and x^m indicates the m-th modality.

Lemma 3.2. *Suppose we have two versions of a sample $x^{(\mathbb{M})}$, i.e., $x^{(\mathbb{T})}$ and $x^{(\mathbb{S})}$, if we can assure $\mathbb{T} \subset \mathbb{S} \subseteq \mathbb{M}$, then, for any class label y, we have $I(y, x^{(\mathbb{T})}) \leq I(y, x^{(\mathbb{S})})$.*

In other words, $x^{(\mathbb{S})}$ is more predictive for the target than $x^{(\mathbb{T})}$ regardless of the label. For a trustworthy multimodal classification model, the confidence of $x^{(\mathbb{T})}$ should not be larger than $x^{(\mathbb{S})}$.

Algorithm 1 Calibrating Multimodal Classifier

```
**Given** dataset $\mathcal{D} = \left\{\{x_i^m\}_{m=1}^M, y_i\right\}_{i=1}^N$, initialized classifier $f$, classification loss criterion $\mathcal{L}^{\mathrm{CL}}$, hyperparameter $\lambda$, and epochs for training the classifier $train\_epochs$.
**for** $e = 1, \ldots, train\_epochs$ **do**
  $\mathbb{S} \leftarrow \mathbb{M}$; $\mathcal{L}^{\mathrm{CL}} \leftarrow \mathcal{L}^{\mathrm{CL}}(x^{(\mathbb{S})})$; $\mathcal{L}^{\mathrm{CML}} \leftarrow 0$;
  **for** $m = 1, \ldots, M-1$ **do**
    Randomly remove a modality of $\mathbb{S}$ and set it as $\mathbb{T}$;
    Compute the classification loss: $\mathcal{L}^{\mathrm{CL}} \leftarrow \mathcal{L}^{\mathrm{CL}} + \mathcal{L}^{\mathrm{CL}}(x^{(\mathbb{T})})$;
    Compute the regularization loss: $\mathcal{L}^{\mathrm{CML}} \leftarrow \mathcal{L}^{\mathrm{CML}} + \max\left(0, \mathrm{Conf}(x^{(\mathbb{T})}) - \mathrm{Conf}(x^{(\mathbb{S})}\right)$;
    $\mathbb{S} \leftarrow \mathbb{T}$;
  **end for**
  Total loss: $\mathcal{L} = \frac{1}{M}\mathcal{L}^{\mathrm{CL}} + \lambda\mathcal{L}^{\mathrm{CML}}$;
  Update the parameters of the classifier $f$ with $\mathcal{L}$;
**end for**
**return** the classifier $f$
```

○ **Why can CML regularization calibrate a model?** CML regularization can guarantee a smaller confidence of $x^{(\mathbb{T})}$ when the model makes a wrong prediction of $x^{(\mathbb{S})}$, which means that CML can alleviate the over-confidence.

Lemma 3.3. *Suppose the CML regularization can achieve a lower* VRR, *i.e.,* $\mathrm{VRR}_{CML} < \mathrm{VRR}_{\mathit{ORIG}}$, *then for the samples that meet* $\mathbb{E}\left(\mathrm{Conf}_{CML}(x^{(\mathbb{S})})\right) = \mathbb{E}\left(\mathrm{Conf}_{\mathit{ORIG}}(x^{(\mathbb{S})})\right)$, *we have* $\mathbb{E}\left(\mathrm{Conf}_{CML}(x^{(\mathbb{T})})\right) \leq \mathbb{E}\left(\mathrm{Conf}_{\mathit{ORIG}}(x^{(\mathbb{T})})\right)$.

From the empirical results, we find $\mathrm{Conf}_{CML}(x^{(\mathbb{S})})$ and $\mathrm{Conf}_{ORIG}(x^{(\mathbb{S})})$ are very similar for most samples, where $\mathrm{Conf}_{ORIG}(\cdot)$ and $\mathrm{Conf}_{CML}(\cdot)$ indicate the confidence estimated by the original (ORIG) model and the model improved by CML regularization respectively. The proof of Lemma 3.3 and empirical results please refer to Appendix B.5.

○ **Why not just penalize the difference in confidence (i.e., minimizing Conf$(x^{(\mathbb{T})})$ − Conf$(x^{(\mathbb{S})})$)?** Forcing the confidence for $x^{(\mathbb{T})}$ to be smaller than the confidence for $x^{(\mathbb{S})}$ regardless of whether the samples violate the Prop. 3.1 will lead to very small confidence for $x^{(\mathbb{T})}$, and adding such a penalty to samples who meet the Prop. 3.1 will lead to a trivial solution (i.e., extremely small confidence when any modality is removed, and the experiments are shown in Appendix B.6). What's more, the model sometimes can still make correct predictions confidently when one modality is removed. A flexible ranking regularization (Eq. 4) makes it more reasonable for the real situation.

Table 1: VRR (%) of test samples (a lower value indicates a better confidence estimation. Type III is shown in Appendix). "✗" indicates the model is not equipped with the proposed regularization ($\lambda = 0$). Performance on Type III please refer to Appendix B.6.

Method	CML	TUANDROMD	YaleB	Handwritten	CUB	Animal
Type I	✗	23.38 ± 1.39	39.15 ± 4.97	17.64 ± 2.31	2.83 ± 1.55	44.39 ± 7.55
	✓	12.58 ± 2.84	15.05 ± 1.12	3.18 ± 0.80	2.17 ± 1.13	29.02 ± 5.43
	Improve	△ 10.80	△ 24.10	△ 14.46	△ 0.66	△ 15.37
Type II	✗	39.17 ± 2.32	20.54 ± 4.26	33.82 ± 5.16	23.17 ± 4.87	12.51 ± 1.50
	✓	8.38 ± 1.31	14.46 ± 2.17	29.99 ± 2.30	20.17 ± 3.05	8.64 ± 0.32
	Improve	△ 30.79	△ 6.08	△ 3.83	△ 3.00	△ 3.87

4. Experiments

4.1. Setup

We deploy the proposed regularization strategy into different types of multimodal classifiers including the imputation-independent method (Type I), the imputation-dependent method (Type II), and the recent state-of-the-art method (Type III). CPM-Nets (Zhang et al., 2019) is a typical imputation-independent algorithm, which can adapt to arbitrary missing patterns without reconstructing the missing modalities. MIWAE (Mattei & Frellsen, 2019) is a imputation-dependent algorithm. The above two methods are well-established models in incomplete multimodal learning. In addition to incomplete multimodal learning methods, we also deploy the regularization into an advanced multimodal classification method (Wu et al., 2022), which is termed Multimodal Transfer Module (MMTM). We approximate the modality removal by feature corruption (e.g., adding strong noise) because MMTM can not make a prediction when one modality is explicitly removed. For a fair comparison, the only difference between whether the model is equipped with CML regularization or not. Please refer to Appendix B.2 for more detailed settings.

Datasets: We evaluate the proposed method on diverse datasets, including data with multimodal data, such as YaleB (Georghiades et al., 2002), Handwritten (Perkins & Theiler, 2003), CUB (Wah et al., 2011), Animal (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015) (which is a dataset under class-imbalanced), TUANDROMD (Borah et al., 2020), NYUD2 (Qi et al., 2017), and SUN-RGBD (Song et al., 2015). It should be pointed out that we also estimate the proposed method on the class-imbalanced dataset. We find that CML can improve the performance when the training data is class-imbalanced since CML calibrates the model regardless of the label while the vanilla model always tends to be under-confidence of the minority classes compared with majority classes. For more detailed analysis please refer to Appendix B.1.

4.2. Questions to be Verified

We conduct diverse experiments to comprehensively investigate the underlying assumption and the proposed method, including:

○ **Can CML regularization improve the confidence estimation of multimodal classifiers?** To validate whether the proposed method improves multimodal classifiers' confidence estimation, we evaluate the confidence estimation of current multimodal classifiers without and with CML regularization, respectively. We conduct experiments of each type of method on seven datasets and evaluate their trustworthiness in terms of VRR (defined in Eq. 2).

○ **Can CML regularization improve robustness?** CML regularization can improve multimodal classifiers' confidence estimation, so a natural question arises - does a better confidence estimation imply better robustness? To verify this, we evaluate the robustness on the complete multimodal data and noisy multimodal data (adding Gaussian noise to some modalities, i.e., zero mean with varying variance ϵ).

○ **Is CML easy to be deployed and not sensitive to hyperparameters?** In order to investigate the key factor that makes the improvement in the proposed method, we evaluate the performance in terms of classification accuracy under different strengths of CML regularization. We conduct experiments on both the original and noised data (i.e., adding noise to one of the modalities during the test). More details are shown in Appendix B.2.

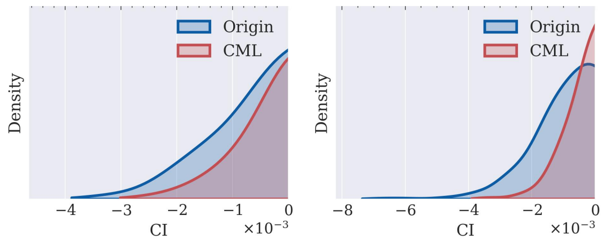
4.3. Results

4.3.1. Confidence Estimation

We evaluate the confidence estimation of current multimodal learning models from a ranking perspective. It is observed that for a large portion of samples the confidence will increase when one modality is removed, while the confidence estimation of the classification models equipped with our proposed CML regularization is significantly improved. We intuitively demonstrate the confidence changing in Fig. 3,

Table 2: Accuracy performance comparison for whether the model is equipped with the CML regularization term (i.e., whether λ is set to 0). The means and standard deviations over five runs are reported.

Method	Dataset	CML	Accuracy (↑)	NLL (↓)	AURC (↓)	E-AURC (↓)
Type I	CUB	✗	87.00 ± 4.36	20.49 ± 0.30	59.44 ± 22.10	49.52 ± 17.35
		✓	88.33 ± 4.05	20.53 ± 0.46	55.94 ± 17.07	47.92 ± 16.89
		Improve	△ 1.33	▽ 0.04	△ 3.50	△ 1.60
	Animal	✗	81.72 ± 2.51	36.87 ± 0.41	82.14 ± 27.20	63.94 ± 22.74
		✓	82.73 ± 1.64	36.87 ± 0.36	71.54 ± 16.03	55.50 ± 13.13
		Improve	△ 1.01	0.00	△ 10.60	△ 8.44
	TUANDROMD	✗	84.66 ± 0.43	6.88 ± 0.00	61.46 ± 6.09	49.00 ± 5.75
		✓	85.20 ± 0.81	6.88 ± 0.00	58.24 ± 5.05	46.64 ± 4.55
		Improve	△ 0.54	0.00	△ 3.22	△ 2.36
Type II	CUB	✗	92.33 ± 1.11	2.33 ± 0.55	10.92 ± 1.94	7.82 ± 1.32
		✓	94.50 ± 1.71	2.24 ± 1.27	9.32 ± 3.91	7.60 ± 3.02
		Improve	△ 2.17	△ 0.09	△ 1.60	△ 0.22
	Animal	✗	86.75 ± 0.33	8.25 ± 3.79	27.62 ± 7.42	18.40 ± 7.27
		✓	87.61 ± 0.50	4.99 ± 0.46	21.26 ± 1.31	13.24 ± 0.92
		Improve	△ 0.86	△ 3.26	△ 6.36	△ 5.16
	TUANDROMD	✗	86.32 ± 0.85	3.26 ± 0.09	43.40 ± 2.65	33.56 ± 2.38
		✓	88.69 ± 0.99	3.21 ± 0.15	38.62 ± 5.44	31.90 ± 4.37
		Improve	△ 2.37	△ 0.02	△ 4.78	△ 1.66
Type III	NYUD2	✗	66.89 ± 0.85	10.03 ± 0.10	140.53 ± 5.66	78.40 ± 5.01
		✓	68.09 ± 0.68	9.83 ± 0.15	137.27 ± 6.94	79.87 ± 6.30
		Improve	△ 1.20	△ 0.20	△ 3.26	▽ 1.47
	SUN-RGBD	✗	62.11 ± 0.31	13.27 ± 0.53	181.00 ± 1.20	97.87 ± 1.48
		✓	62.78 ± 0.32	13.25 ± 0.46	174.90 ± 1.50	95.00 ± 1.00
		Improve	△ 0.67	△ 0.05	△ 6.10	△ 2.87



(a) Animal

(b) Tuandromd

Figure 3: Confidence estimation when one modality is removed, where "CI" is defined in Eq. 1.

and the quantitative results are shown in Tab. 1. According to Fig. 3, we show the confidence estimation of CPM-Nets, where "Original" and "CML" indicate the model is without and with the proposed CML regularization respectively. According to Fig. 3, it is observed that the confidence without CML regularization may increase when one modality is removed, which indicates that the model fails to take all modalities into account fairly when making predictions. This will lead to unpromising robustness and generalization, which clearly verifies the main assumption in Sec. 4.3.2.

4.3.2. CML Regularization Improves Robustness

In this subsection, we evaluate the performance on the complete multimodal data, where the training/test data is divided as previous work (Zhang et al., 2019). From Tab. 2, the classification models equipped with CML regularization consistently outperform their counterparts (i.e., the original classification models) validating the rationality of CML principle. It is worth noting that Type III exhibits a significant improvement, while the improvement in Type I and Type II is relatively minor compared to the standard deviation. The high variance can be attributed to the baseline models themselves. To avoid the influence of empirical contingency, we report the means and standard deviations over 5 or 10 runs in our paper. Furthermore, we distinguish the marks in the table based on the significance of the improvement, with a lighter color indicating a relatively minor improvement compared to the standard deviation. Results on more

Table 3: Accuracy performance comparison when some of the modalities is corrupted with Gaussian noise (i.e., zero mean with varying variance ϵ).

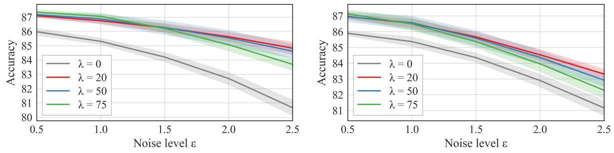
Dataset	Noise on	CML	$\epsilon = \mathbf{0.1}$	$\epsilon = \mathbf{0.2}$	$\epsilon = \mathbf{0.3}$	$\epsilon = \mathbf{0.5}$
CUB	{1}	✗	84.72 ± 3.32	82.22 ± 4.53	79.72 ± 4.43	71.17 ± 9.14
		✓	85.83 ± 2.72	85.00 ± 3.50	84.17 ± 4.08	81.11 ± 4.37
		Improve	△ 1.11	△ 2.78	△ 4.45	△ 9.94
	{2}	✗	84.44 ± 2.75	83.89 ± 3.22	83.61 ± 2.83	83.61 ± 3.87
		✓	85.83 ± 3.40	85.28 ± 2.75	85.28 ± 1.97	85.00 ± 1.80
		Improve	△ 1.39	△ 1.39	△ 1.67	△ 1.39
	{1, 2}	✗	85.00 ± 3.12	82.78 ± 3.98	80.00 ± 4.46	72.50 ± 11.14
		✓	85.83 ± 2.72	85.84 ± 3.12	85.83 ± 4.25	81.39 ± 6.43
		Improve	△ 0.83	△ 3.06	△ 5.83	△ 8.89
Animal	{1}	✗	80.78 ± 2.79	80.96 ± 2.78	80.85 ± 2.80	80.68 ± 2.93
		✓	82.03 ± 1.91	82.37 ± 2.09	82.55 ± 2.24	82.30 ± 2.40
		Improve	△ 1.25	△ 1.41	△ 1.70	△ 1.62
	{2}	✗	80.70 ± 2.45	79.81 ± 3.14	77.34 ± 4.80	68.52 ± 9.68
		✓	82.07 ± 1.57	81.23 ± 2.32	78.93 ± 3.65	72.39 ± 8.35
		Improve	△ 1.37	△ 1.42	△ 1.59	△ 3.87
	{1, 2}	✗	80.87 ± 2.55	79.97 ± 3.12	77.11 ± 5.86	65.08 ± 12.75
		✓	82.14 ± 1.76	81.95 ± 2.65	79.63 ± 5.28	72.46 ± 11.39
		Improve	△ 1.27	△ 1.98	△ 2.52	△ 7.38

datasets are shown in Appendix B.4.

Significantly improving the accuracy on real-world data without additional techniques or more advanced architectures can be challenging as the benchmark datasets have already achieved good performance in terms of accuracy. However, we observed that the models equipped with CML regularization are more robust to noise, particularly when the noise is heavy. Specifically, we find that CML regularization can improve the robustness of imperfect data, such as noise. We evaluate the models in terms of the accuracy in the test under Gaussian noise (i.e., zero mean and varying variance ϵ), and "Noise On" indicates which modality is noised (e.g., $\{1\}$ indicates the first modality is noised). We report the performance on the challenging datasets (CUB and Animal) in the main text (Tab. 3) and more results are in Appendix B.3. We can find that the models equipped with CML regularization are more robust to noise, especially when the noise is much heavier.

4.3.3. Performance under Different Strengths of CML Regularization

In this subsection, we report the accuracy under different strengths of regularization (where "$\lambda = 0$" indicates the model is not equipped with the proposed CML regularization). We also add Gaussian noise (i.e., zero mean and varying variance ϵ) to one of the modalities on CUB, and it is clear that the model with CML regularization is more robust to the potential noise.



(a) Noise on the first modality (b) Noise on the second modality

Figure 4: Accuracy estimation where one of the modalities is corrupted with noise.

As shown in Fig. 4, it is observed that CML regularization can promote accuracy on the noisy data. The potential reason is that the CML regularization enforces the reasonable confidence estimation and thus prohibits the model from being over-confident on the low-quality modality, where the low-quality modality usually tends to result in a wrong decision. Moreover, according to Fig. 4, the proposed regularization is not sensitive to the hyperparameter λ, where promising performance could be expected with a mild regularization strength. In other words, the proposed regularization is not sensitive to hyperparameters and CML is easy to be deployed into a wide spectrum of multimodal models.

5. Conclusion

In this work, we reveal a novel issue widely existing in multimodal learning through extensive empirical studies. We observe that the confidence estimations of current multimodal learning algorithms are typically unreliable, and tend to rely on some partial modalities. This further results in the non-robustness of learned models against modality corruption. Concretely, existing multimodal classifiers tend to be overconfident based on some modalities, and ignore the valuable evidence from other modalities even those might be critical to make the decision. To solve this problem, we introduce a novel regularization technique to calibrate the confidence estimation, which forces model to estimate a calibrated predictive confidence. This technique can be naturally deployed into existing multimodal learning methods without modifying the main training process. We conduct comprehensive experiments which demonstrate the superiority of our method in classification in terms of both accuracy and calibration. The proposed method is the first attempt to calibrate the relationship between confidence and the number of modalities used in multimodal learning. This research is an inspirational topic which could benefit the multimodal learning community. In current implementation, we employ sampling to construct constraint. Although it is widely used and effective in machine learning, we will focus on more principled approximation strategies in the future.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China (Grant No. 61976151), the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A1b0045), and A*STAR Central Research Fund. We gratefully acknowledge the support of CAAI-Huawei MindSpore Open Fund[1]. The project was finished during the internship in AI Lab, Tencent.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., and Parikh, D. Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, 2015.

Bagher Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

Bai, Y., Mei, S., Wang, H., and Xiong, C. Understanding the under-coverage bias in uncertainty estimation. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18307–18319. Curran Associates, Inc., 2021.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100, 1998.

Borah, P., Bhattacharyya, D., and Kalita, J. Malware dataset generation and evaluation. In *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, pp. 1–6. IEEE, 2020.

Burgin, M. The essence of information: Paradoxes, contradictions, and solutions. In *Electronic Conference on Foundations of Information Science: The nature of information: Conceptions, misconceptions, and paradoxes (FIS 2002). Retrieved September*, volume 13, pp. 2013. Citeseer, 2002.

Chau, S. L., Ton, J.-F., González, J., Teh, Y., and Sejdinovic, D. Bayesimp: Uncertainty quantification for causal data fusion. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3466–3477. Curran Associates, Inc., 2021.

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 10971–10984. Curran Associates, Inc., 2021.

Cobb, A. D. and Jalaian, B. Scaling hamiltonian monte carlo inference for bayesian neural networks with symmetric splitting. In *Uncertainty in Artificial Intelligence*, pp. 675–685. PMLR, 2021.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. Addressing failure prediction by learning model confidence. In *NeurIPS*, 2019.

Denker, J. and LeCun, Y. Transforming neural-net output levels to probability distributions. *Advances in neural information processing systems*, 3, 1990.

Foong, A., Burt, D., Li, Y., and Turner, R. On the expressiveness of approximate inference in bayesian neural networks. *NeurIPS*, 33:15897–15908, 2020.

[1] https://www.mindspore.cn/

Galil, I. and El-Yaniv, R. Disrupting deep uncertainty estimation without harming accuracy. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 21285–21296. Curran Associates, Inc., 2021.

Georghiades, A. S., Belhumeur, P. N., and Kriegman, D. J. From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 23 (6):643–660, 2002.

Guo, C., Pleiss, G., Yu, S., and Weinberger, K. Q. On calibration of modern neural networks. In *ICML*, 2017.

Hafner, D., Tran, D., Lillicrap, T. P., Irpan, A., and Davidson, J. Noise contrastive priors for functional uncertainty. In *UAI*, 2019.

Han, X., Wang, S., Su, C., Huang, Q., and Tian, Q. Greedy gradient ensemble for robust visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1584–1593, 2021.

Havasi, M., Jenatton, R., Fort, S., Liu, J. Z., Snoek, J., Lakshminarayanan, B., Dai, A. M., and Tran, D. Training independent subnetworks for robust prediction. *arXiv preprint arXiv:2010.06610*, 2020.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

Jeong, J., Park, S., Kim, M., Lee, H.-C., Kim, D.-G., and Shin, J. Smoothmix: Training confidence-calibrated smoothed classifiers for certified robustness. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30153–30168. Curran Associates, Inc., 2021.

Joze, H. R. V., Shaban, A., Iuzzolino, M. L., and Koishida, K. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13289–13299, 2020.

Karaletsos, T. and Bui, T. D. Hierarchical gaussian process priors for bayesian neural network weights. *NeurIPS*, 33: 17141–17152, 2020.

Karandikar, A., Cain, N., Tran, D., Lakshminarayanan, B., Shlens, J., Mozer, M. C., and Roelofs, B. Soft calibration objectives for neural networks. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29768–29779. Curran Associates, Inc., 2021.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Khodayari, A., Ghaffari, A., Ameli, S., and Flahatgar, J. A historical review on lateral and longitudinal control of autonomous vehicle motions. In *International Conference on Mechanical & Electrical Technology*, 2010.

Kishi, R. M., Trojahn, T. H., and Goularte, R. Correlation based feature fusion for the temporal video scene segmentation task. *Multimedia Tools & Applications*, 78(11): 15623–15646, 2019.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, volume 25, 2012.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

Le, Q. and Mikolov, T. Distributed representations of sentences and documents. In *ICML*, pp. 1188–1196. PMLR, 2014.

Lee, C. and van der Schaar, M. A variational information bottleneck approach to multi-omics data integration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1513–1521. PMLR, 2021.

Luo, M., Chen, F., Hu, D., Zhang, Y., Liang, J., and Feng, J. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5972–5984. Curran Associates, Inc., 2021.

MacKay, D. J. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

Mattei, P.-A. and Frellsen, J. Miwae: Deep generative modelling and imputation of incomplete data sets. In *ICML*, pp. 4413–4423. PMLR, 2019.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15682–15694. Curran Associates, Inc., 2021.

Molchanov, D., Ashukha, A., and Vetrov, D. P. Variational dropout sparsifies deep neural networks. In *ICML*, 2017.

Moon, J., Kim, J., Shin, Y., and Hwang, S. Confidence-aware learning for deep neural networks. In *ICML*, 2020.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., and Dokania, P. Calibrating deep neural networks using focal loss. In *NeurIPS*, 2020.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *NeurIPS*, 2019.

Neal, R. M. *Bayesian learning for neural networks*. Springer Science & Business Media, 2012.

Ning, Q., Dong, W., Li, X., Wu, J., and Shi, G. Uncertainty-driven loss for single image super-resolution. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 16398–16409. Curran Associates, Inc., 2021.

Pan, T.-Y., Zhang, C., Li, Y., Hu, H., Xuan, D., Changpinyo, S., Gong, B., and Chao, W.-L. On model calibration for long-tailed object detection and instance segmentation. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2529–2542. Curran Associates, Inc., 2021.

Pérez-Rúa, J.-M., Vielzeuf, V., Pateux, S., Baccouche, M., and Jurie, F. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6966–6975, 2019.

Perkins, S. and Theiler, J. Online feature selection using grafting. In *ICML*, 2003.

Perrin, R. J., Fagan, A. M., and Holtzman, D. M. Multimodal techniques for diagnosis and prognosis of alzheimer's disease. *Nature*, 461(7266):916–922, 2009.

Qaddoum, K. and Hines, E. L. Reliable yield prediction with regression neural networks. In *WSEAS international conference on systems theory and scientific computation*, 2012.

Qi, X., Liao, R., Jia, J., Fidler, S., and Urtasun, R. 3d graph neural networks for rgbd semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5199–5208, 2017.

Qin, Y., Wang, X., Beutel, A., and Chi, E. Improving calibration through the relationship with adversarial robustness. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 14358–14369. Curran Associates, Inc., 2021.

Rahaman, R. and thiery, a. Uncertainty quantification and deep ensembles. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 20063–20075. Curran Associates, Inc., 2021.

Ritter, H., Kukla, M., Zhang, C., and Li, Y. Sparse uncertainty representation in deep learning with inducing weights. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6515–6528. Curran Associates, Inc., 2021.

Sahoo, R., Zhao, S., Chen, A., and Ermon, S. Reliable decisions with threshold calibration. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1831–1844. Curran Associates, Inc., 2021.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

Singh, A., Kempe, D., and Joachims, T. Fairness in ranking under uncertainty. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11896–11908. Curran Associates, Inc., 2021.

Slack, D., Hilgard, A., Singh, S., and Lakkaraju, H. Reliable post hoc explanations: Modeling uncertainty in explainability. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9391–9404. Curran Associates, Inc., 2021.

Song, S., Lichtenberg, S. P., and Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

Soni, J. and Goodman, R. *A mind at play: how Claude Shannon invented the information age*. Simon and Schuster, 2017.

Stadler, M., Charpentier, B., Geisler, S., Zügner, D., and Günnemann, S. Graph posterior network: Bayesian predictive uncertainty for node classification. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18033–18048. Curran Associates, Inc., 2021.

Sun, Y., Mai, S., and Hu, H. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters*, 28:1650–1654, 2021.

Tian, J., Yung, D., Hsu, Y.-C., and Kira, Z. A geometric perspective towards neural calibration via sensitivity decomposition. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26358–26369. Curran Associates, Inc., 2021.

Toneva, M., Sordoni, A., Combes, R. T. d., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Upadhyay, U., Chen, Y., and Akata, Z. Robustness via uncertainty-aware cycle consistency. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28261–28273. Curran Associates, Inc., 2021.

van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. Uncertainty estimation using a single deep deterministic neural network. In *ICML*, 2020.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Wald, Y., Feder, A., Greenfeld, D., and Shalit, U. On calibration and out-of-domain generalization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 2215–2227. Curran Associates, Inc., 2021.

Wang, W., Tran, D., and Feiszli, M. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12695–12705, 2020.

Wang, Y. and Zou, S. Online robust reinforcement learning with model uncertainty. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7193–7206. Curran Associates, Inc., 2021.

Wang, Y., Shen, Y., Liu, Z., Liang, P. P., Zadeh, A., and Morency, L.-P. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, 2019.

Wu, M. and Goodman, N. Multimodal generative models for scalable weakly-supervised learning. *NeurIPS*, 31, 2018.

Wu, N., Jastrzębski, S., Cho, K., and Geras, K. J. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, 2022.

Xiong, R., Chen, Y., Pang, L., Cheng, X., Ma, Z.-M., and Lan, Y. Uncertainty calibration for ensemble-based debiasing methods. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13657–13669. Curran Associates, Inc., 2021.

Xu, Z., Chai, Z., and Yuan, C. Towards calibrated model for long-tailed visual recognition from prior perspective. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7139–7152. Curran Associates, Inc., 2021.

Zaidi, S., Zela, A., Elsken, T., Holmes, C. C., Hutter, F., and Teh, Y. Neural ensemble search for uncertainty estimation and dataset shift. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7898–7911. Curran Associates, Inc., 2021.

Zhang, C., Han, Z., cui, y., Fu, H., Zhou, J. T., and Hu, Q. Cpm-nets: Cross partial multi-view networks. In *NeurIPS*, volume 32, 2019.

Zhang, Y., Wang, C., and Deng, W. Relative uncertainty learning for facial expression recognition. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17616–17627. Curran Associates, Inc., 2021.

Zhao, S., Kim, M., Sahoo, R., Ma, T., and Ermon, S. Calibrating predictions to decisions: A novel approach to multi-class calibration. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22313–22324. Curran Associates, Inc., 2021.

A. How to Make Ranking Pairs

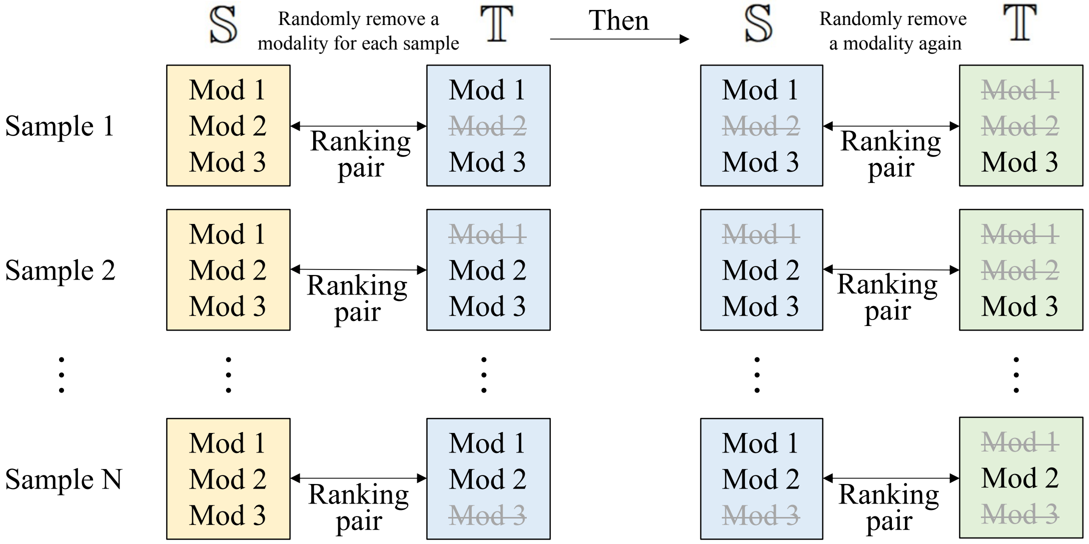


Figure 5: Illustration of generating $\mathbb{S}$ and $\mathbb{T}$.

To compute this score in practice, following the prior methods (Moon et al., 2020; Toneva et al., 2018) we initialize $\mathbb{S}$ as the complete modalities, and obtain $\mathbb{T}$ by randomly removing a modality from $\mathbb{S}$. Then $\mathbb{T}$ is regarded as $\mathbb{S}$ for another confidence ranking pair and we repeat this process until there is only one modality remained in $\mathbb{T}$.

B. Experiments Details

B.1. Dataset Details

We evaluate the proposed method on diverse datasets, including data with multiple modalities and multiple types of features. ○ **YaleB**: Similar to previous work (Georghiades et al., 2002), we also use a subset of this face image dataset, which contains 650 facial images, 10 classes and 3 different types of features. ○ **Handwritten** (Perkins & Theiler, 2003): This is a database of handwritten digits which contains $2,000$ images, 10 classes, 6 types of features. ○ **CUB** (Wah et al., 2011): Following CPM-Nets (Zhang et al., 2019), we use a subset of this dataset, which contains first 10 classes of original dataset and 2 modalities (deep visual feature and text feature) are obtained by GoogleNet and doc2vec (Le & Mikolov, 2014). ○ **Animal**: This dataset contains $10,158$ images, 50 classes, and 2 types of features (deep visual feature from DECAF (Krizhevsky et al., 2012) and VGG19 (Simonyan & Zisserman, 2015)). ○ **TUANDROMD** (Borah et al., 2020): The dataset contains $4,465$ instances, 2 classes and 2 types of modalities.

B.2. Experiment Setting

Type-I: For CPM-Nets and the first five datasets(i.e.,YaleB, Handwritten, CUB and Animal), we follow the author's implementation (Zhang et al., 2019): the dimensionality of latent representation is 150. Parameter lambda for cub/animal/hand-written/yaleB/tuandromd is set as 5/45/45/10/5. The dimensionalities of input, hidden layers are 128 and 300. We use Adam optimizer to train all CPM-Nets models with the learning rate of 10^{-2} and no additional regularization term. For Tuandromd dataset, we tune the dimensionality of latent representation to 512. The dimensionalities of input and hidden layers are both 512. We use Adam optimizer to train CPM-Net with L2-regularization term. **Type-II**: For MIWAE, we train the encoder, decoder and classifier respectively. The number of hidden units of them is all 128. Parameter lambda for cub/animal/hand-written/yaleB/tuandromd are set as 15/25/10/35/75 for best performance. The dimensionalities of the latent space are 64. We use Adam optimizer to train the encoder and decoder with a learning rate of 10^{-2}. Then we train the encoder, decoder and classifier altogether for another with a learning rate of 10^{-3}. As same as prior work (Corbière et al., 2019), we evaluated the performance according to Accuracy (%), NLL (10^{-1}), AURC (10^{-3}), and E-AURC (10^{-3}).

Table 4: Accuracy performance comparison when some of the modalities is blurred (Type I).

Dataset	Noise on	CML	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$	$\epsilon = 0.4$	$\epsilon = 0.5$
YaleB	{1}	✗	97.43 ± 1.58	96.92 ± 1.88	96.41 ± 2.20	94.10 ± 1.31	92.82 ± 1.31
		✓	**98.46 ± 1.09**	**98.20 ± 1.31**	**96.15 ± 1.88**	**94.62 ± 1.88**	**93.59 ± 1.30**
	{2}	✗	95.13 ± 0.72	94.10 ± 1.31	92.57 ± 0.73	92.05 ± 1.45	91.54 ± 1.66
		✓	**96.92 ± 1.26**	**95.90 ± 2.02**	**94.61 ± 2.88**	**93.33 ± 2.54**	**93.08 ± 3.14**
	{3}	✗	94.87 ± 0.96	94.87 ± 0.96	94.10 ± 0.96	92.82 ± 1.81	92.05 ± 1.31
		✓	**96.92 ± 1.88**	**97.18 ± 1.92**	**96.15 ± 1.88**	**94.87 ± 2.54**	**94.36 ± 2.02**
	{1, 2}	✗	96.67 ± 2.61	95.13 ± 3.46	91.28 ± 2.83	88.72 ± 3.10	86.41 ± 3.10
		✓	**97.69 ± 0.63**	**95.39 ± 2.26**	**92.56 ± 2.02**	**89.72 ± 2.21**	**86.66 ± 1.81**
	{1, 3}	✗	97.43 ± 0.96	97.69 ± 1.66	97.43 ± 1.81	97.18 ± 2.20	96.15 ± 2.26
		✓	**98.46 ± 1.09**	**98.46 ± 1.26**	**98.46 ± 1.66**	**96.92 ± 1.88**	**96.67 ± 2.20**
	{2, 3}	✗	94.62 ± 1.08	93.85 ± 1.25	90.26 ± 2.54	87.95 ± 2.83	86.67 ± 2.38
		✓	**96.41 ± 1.81**	**95.64 ± 1.92**	**93.84 ± 3.32**	**91.28 ± 3.10**	**89.49 ± 3.16**
	{1, 2, 3}	✗	96.15 ± 1.88	96.41 ± 3.16	93.85 ± 4.40	87.69 ± 8.21	84.10 ± 10.32
		✓	**97.43 ± 1.81**	**97.43 ± 1.92**	**93.85 ± 4.40**	**87.69 ± 7.61**	**82.56 ± 9.26**
Hand-written	{1}	✗	97.18 ± 1.92	95.38 ± 1.25	93.34 ± 1.31	92.57 ± 1.58	91.28 ± 1.31
		✓	**98.46 ± 1.26**	**95.90 ± 1.92**	**93.85 ± 1.88**	**93.08 ± 1.66**	**92.31 ± 0.63**
	{2}	✗	88.46 ± 1.66	87.18 ± 1.31	86.92 ± 1.09	86.92 ± 1.09	86.92 ± 1.09
		✓	**90.77 ± 3.33**	**90.26 ± 3.57**	**89.75 ± 3.85**	**89.75 ± 3.84**	**89.75 ± 3.84**
	{3}	✗	85.90 ± 1.92	85.13 ± 1.81	84.87 ± 1.45	84.62 ± 1.66	84.62 ± 1.66
		✓	**88.97 ± 2.54**	**88.21 ± 2.61**	**87.69 ± 2.74**	**87.69 ± 3.32**	**87.44 ± 3.10**
	{1, 2}	✗	88.97 ± 3.68	83.08 ± 3.50	78.97 ± 1.92	77.69 ± 2.74	75.90 ± 3.57
		✓	**88.97 ± 4.04**	**83.59 ± 2.97**	**80.51 ± 3.46**	**77.18 ± 4.28**	**74.10 ± 3.84**
	{1, 3}	✗	91.54 ± 1.09	91.28 ± 3.16	88.97 ± 5.41	87.43 ± 5.83	85.64 ± 6.42
		✓	**93.59 ± 2.38**	**91.79 ± 3.68**	**88.97 ± 4.04**	**86.93 ± 4.99**	**85.39 ± 4.91**
	{2, 3}	✗	63.59 ± 8.00	**59.74 ± 7.00**	**57.69 ± 5.99**	**56.67 ± 5.94**	**55.90 ± 5.49**
		✓	**64.36 ± 7.49**	58.46 ± 6.37	56.67 ± 6.10	55.64 ± 6.04	54.87 ± 6.29
	{1, 2, 3}	✗	54.87 ± 10.68	**37.95 ± 6.92**	**29.48 ± 4.76**	**24.36 ± 4.04**	**22.31 ± 4.12**
		✓	**57.18 ± 11.41**	35.64 ± 4.80	26.67 ± 2.54	22.82 ± 2.54	20.77 ± 1.09
TUAND-ROMD	{1}	✗	84.77 ± 0.55	80.47 ± 0.99	76.53 ± 1.11	72.65 ± 0.76	70.17 ± 0.66
		✓	**86.50 ± 0.59**	**82.46 ± 0.77**	**78.30 ± 1.18**	**74.92 ± 1.39**	**72.45 ± 1.33**
	{2}	✗	86.56 ± 0.27	85.71 ± 0.48	84.14 ± 0.58	82.35 ± 0.86	80.85 ± 1.05
		✓	**88.87 ± 0.22**	**88.74 ± 0.28**	**88.58 ± 0.63**	**88.15 ± 0.65**	**87.93 ± 0.67**
	{1, 2}	✗	84.88 ± 1.19	80.72 ± 1.02	76.60 ± 0.75	73.15 ± 1.10	70.35 ± 1.25
		✓	**87.41 ± 3.40**	**82.78 ± 1.14**	**79.28 ± 1.00**	**76.30 ± 1.11**	**73.82 ± 1.35**

B.3. Robustness Evaluation

We evaluate models in terms of accuracy under Gaussian noise (i.e., zero mean and varying variance ϵ), and "Noise On" indicates which modality is noised (e.g., $\{1\}$ indicates the first modality is noised). In addition to the performance on the challenging datasets (CUB and Animal) in the main text (Table 3), we show more other results (Table 4 5). It is clear that the models equipped with CML are more robust to noise, especially when the noise is much heavier.

Table 5: Accuracy performance comparison when some of the modalities is blurred (Type II).

Dataset	Noise Noise on	CML	$\epsilon = \mathbf{0.5}$	$\epsilon = \mathbf{1.0}$	$\epsilon = \mathbf{1.5}$	$\epsilon = \mathbf{2.0}$	$\epsilon = \mathbf{2.5}$
YaleB	{1}	✗	95.90 ± 2.54	94.87 ± 3.22	93.85 ± 2.88	93.59 ± 3.16	93.59 ± 3.16
		✓	**97.43 ± 1.31**	**96.15 ± 2.51**	**95.13 ± 2.97**	**94.36 ± 2.97**	**93.85 ± 3.46**
	{2}	✗	96.15 ± 2.26	93.33 ± 3.22	91.03 ± 2.62	90.26 ± 2.02	89.23 ± 2.18
		✓	**97.69 ± 1.26**	**96.67 ± 1.58**	**94.10 ± 2.20**	**92.82 ± 2.83**	**92.05 ± 2.02**
	{3}	✗	98.72 ± 0.36	96.92 ± 1.26	96.15 ± 0.63	96.15 ± 0.63	95.90 ± 0.96
		✓	**98.72 ± 0.73**	**97.69 ± 1.09**	**97.43 ± 0.96**	**97.18 ± 1.31**	**96.67 ± 1.58**
	{1, 2}	✗	95.64 ± 2.83	91.02 ± 3.46	88.46 ± 4.53	87.18 ± 3.46	85.90 ± 4.09
		✓	**96.66 ± 1.31**	**93.59 ± 2.38**	**90.51 ± 2.97**	**86.67 ± 3.46**	**84.62 ± 3.26**
	{1, 3}	✗	98.46 ± 0.63	98.46 ± 1.66	97.69 ± 1.66	97.43 ± 1.45	97.18 ± 1.31
		✓	**98.20 ± 0.73**	**97.95 ± 1.92**	**97.69 ± 1.66**	**98.20 ± 1.58**	**97.69 ± 1.66**
	{2, 3}	✗	97.43 ± 0.36	95.89 ± 0.36	95.38 ± 0.62	94.62 ± 0.62	92.82 ± 0.73
		✓	**98.72 ± 0.36**	**97.69 ± 1.09**	**96.66 ± 0.73**	**95.38 ± 0.62**	**94.61 ± 1.66**
	{1, 2, 3}	✗	97.69 ± 0.63	95.64 ± 0.36	93.08 ± 1.09	89.23 ± 1.66	82.31 ± 1.26
		✓	**98.46 ± 0.63**	**97.18 ± 1.31**	**95.64 ± 0.96**	**92.56 ± 2.54**	**88.46 ± 2.27**
CUB	{1}	✗	91.11 ± 1.04	86.94 ± 2.83	83.61 ± 3.93	80.83 ± 4.14	79.17 ± 3.79
		✓	**93.33 ± 1.80**	**90.83 ± 2.45**	**87.50 ± 3.60**	**85.56 ± 4.38**	**81.11 ± 4.53**
	{2}	✗	91.11 ± 0.40	91.95 ± 0.39	91.11 ± 0.40	89.72 ± 0.39	88.61 ± 0.79
		✓	**93.61 ± 1.04**	**92.78 ± 1.04**	**92.50 ± 1.80**	**91.67 ± 2.96**	**91.39 ± 3.22**
	{1, 2}	✗	92.78 ± 1.97	88.61 ± 1.42	85.83 ± 1.80	79.72 ± 2.83	74.17 ± 4.46
		✓	**94.72 ± 2.19**	**92.22 ± 3.75**	**90.00 ± 4.46**	**86.11 ± 4.10**	**79.17 ± 4.91**
Animal	{1}	✗	86.61 ± 0.20	85.81 ± 0.36	84.82 ± 1.02	83.77 ± 1.29	82.16 ± 2.32
		✓	**87.20 ± 0.18**	**87.01 ± 0.18**	**86.60 ± 0.20**	**86.03 ± 0.04**	**85.42 ± 0.29**
	{2}	✗	86.33 ± 0.54	85.62 ± 0.61	84.84 ± 0.95	83.04 ± 1.24	81.34 ± 1.73
		✓	**87.04 ± 0.08**	**86.64 ± 0.26**	**85.95 ± 0.42**	**84.78 ± 0.17**	**82.71 ± 0.24**
	{1, 2}	✗	86.01 ± 0.17	84.80 ± 0.81	83.17 ± 1.65	80.92 ± 2.77	77.42 ± 4.14
		✓	**87.04 ± 0.42**	**86.50 ± 0.15**	**85.38 ± 0.34**	**83.84 ± 0.65**	**81.67 ± 0.75**
TUAND-ROMD	{1}	✗	81.14 ± 0.70	78.21 ± 0.92	75.39 ± 1.09	73.21 ± 1.46	71.71 ± 1.26
		✓	**81.99 ± 1.99**	**78.79 ± 2.42**	**76.37 ± 2.57**	**74.36 ± 2.63**	**73.19 ± 2.60**
	{2}	✗	84.19 ± 0.82	84.43 ± 0.48	84.46 ± 0.35	84.32 ± 0.45	84.21 ± 0.44
		✓	**84.88 ± 1.62**	**84.73 ± 1.89**	**84.84 ± 1.76**	**84.39 ± 0.89**	**84.97 ± 1.52**
	{1, 2}	✗	83.56 ± 1.23	80.85 ± 1.30	77.85 ± 1.53	75.90 ± 2.07	74.08 ± 2.22
		✓	**83.99 ± 1.87**	**81.48 ± 2.30**	**78.50 ± 2.30**	**76.73 ± 2.19**	**75.23 ± 2.20**

B.4. Additional Results for Robustness Estimation

Limited by space, we show the performance of model equipped with CML on YaleB and Handwritten. From Table 6, the classification models equipped with CML consistently outperforms their counterpart validating the rationality of CML principle.

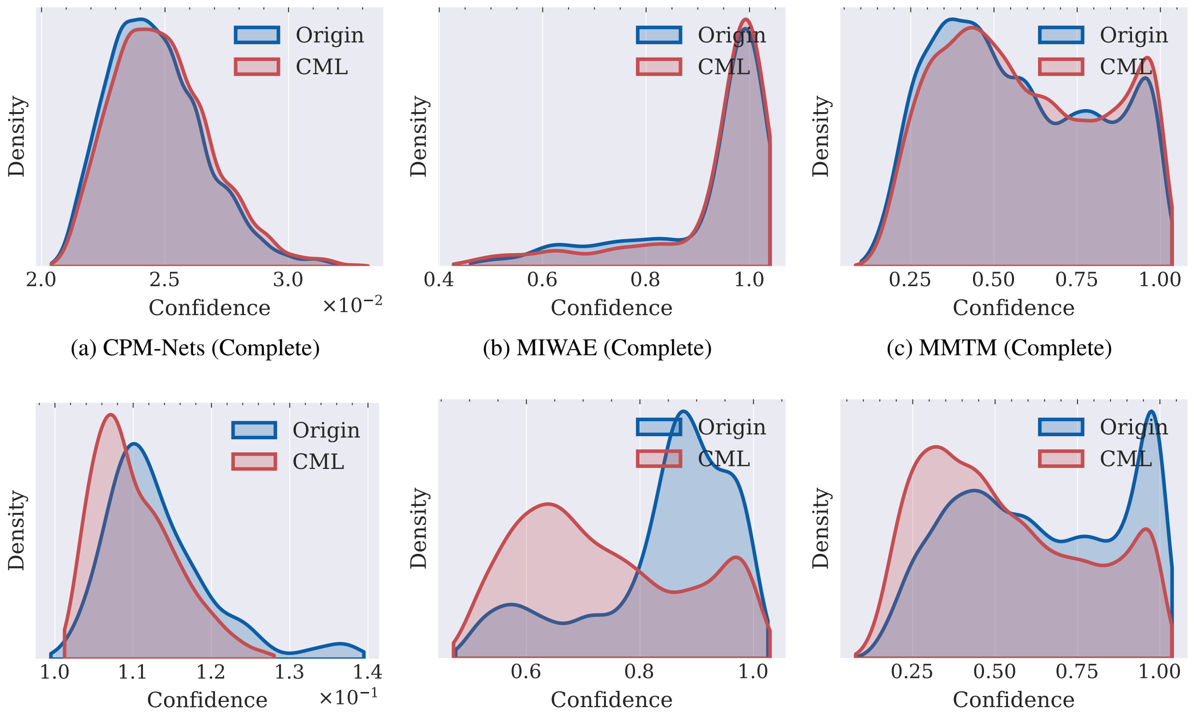
B.5. Confidence Estimation for Complete Inputs

We show the confidence estimation for complete inputs, as shown in Fig. 6, we can find that the confidence estimation of original model and CML model are very similar. To prevent the model from being over-confident when model predicts a wrong prediction, the regularization will not be added when prediction of complete input is wrong. From the bottom figures, we can find CML regularization alleviates the problem that model increases the confidence when one modality is removed.

Proof of Lemma 3.3: if we have $\text{VRR}_{CML} < \text{VRR}_{\textbf{ORIG}}$, then we have $\mathbb{E}\left(\text{Conf}_{CML}(x^{(\mathbb{T})})\right) - \mathbb{E}\left(\text{Conf}_{CML}(x^{(\mathbb{S})})\right) \le$

Table 6: Accuracy performance comparison for whether the model is equipped with the cma regularization term on additional dataset (i.e., whether λ is set to 0).

Method	Dataset	CML	Accuracy (↑)	NLL (↓)	AURC (↓)	E-AURC (↓)
Type I	YaleB	✗	95.84 ± 0.78	21.98 ± 0.05	3.00 ± 1.38	2.08 ± 1.37
		✓	97.69 ± 1.09	21.98 ± 0.05	1.46 ± 1.51	1.12 ± 1.32
		Improve	△ 1.85	0.00	△ 1.54	△ 0.96
	Hand-written	✗	89.00 ± 3.64	20.30 ± 0.25	35.83 ± 20.43	28.80 ± 15.49
		✓	93.60 ± 0.60	20.06 ± 0.11	11.00 ± 6.17	8.90 ± 5.80
		Improve	△ 4.60	△ 0.14	△ 14.83	△ 19.90
Type II	YaleB	✗	95.69 ± 2.10	1.80 ± 0.71	5.50 ± 2.86	4.32 ± 2.32
		✓	97.84 ± 0.58	1.11 ± 0.49	5.02 ± 6.39	4.76 ± 6.26
		Improve	△ 2.15	△ 0.69	△ 0.48	▽ 0.44
	Hand-written	✗	98.40 ± 0.64	0.49 ± 0.12	0.32 ± 0.16	0.16 ± 0.12
		✓	99.05 ± 0.19	0.50 ± 0.10	0.18 ± 0.07	0.14 ± 0.08
		Improve	△ 0.65	0.00	△ 0.14	△ 0.02



(a) CPM-Nets (Complete) (b) MIWAE (Complete) (c) MMTM (Complete)

(d) CPM-Nets (Removed) (e) MIWAE (Removed) (f) MMTM (Removed)

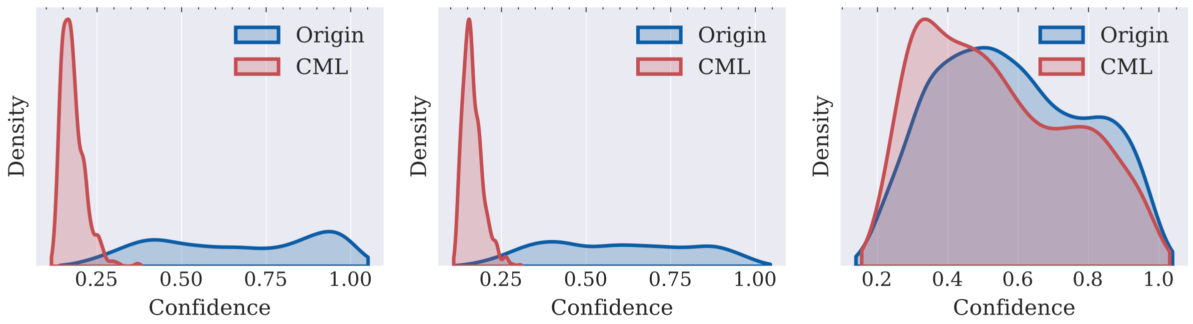
Figure 6: Confidence estimation on complete inputs. We estimate the confidence on complete inputs (top) and the confidence when one modality is removed (bottom). We can find CML regularization keeps the confidence estimation on complete input but alleviate the over-confidence when one modality is removed, which indicates the proposed method calibrates the multimodal model by rethinking the relationship between the modalities.

$\mathbb{E}\left(\text{Conf}_{\text{ORIG}}(x^{(\mathbb{T})})\right) - \mathbb{E}\left(\text{Conf}_{\text{ORIG}}(x^{(\mathbb{S})})\right)$, then we have:

$$\begin{aligned} &\mathbb{E}\left(\text{Conf}_{CML}(x^{(\mathbb{T})})\right) \leq \mathbb{E}\left(\text{Conf}_{\text{ORIG}}(x^{(\mathbb{T})})\right), \\ &\text{subject to: } \mathbb{E}\left(\text{Conf}_{CML}(x^{(\mathbb{T})})\right) = \mathbb{E}\left(\text{Conf}_{ORIG}(x^{(\mathbb{T})})\right) \end{aligned} \tag{7}$$

During the train stage, we evaluate the confidence difference between the $\mathbb{E}\left(\text{Conf}_{CML}(x^{(\mathbb{T})})\right)$ and $\mathbb{E}\left(\text{Conf}_{ORIG}(x^{(\mathbb{T})})\right)$, i.e., $\mathbb{E}\left(\left|\text{Conf}_{CML}(x^{(\mathbb{T})}) - \text{Conf}_{ORIG}(x^{(\mathbb{T})})\right|\right)$. We find the confidence difference between the $\mathbb{E}\left(\text{Conf}_{CML}(x^{(\mathbb{T})})\right)$ and $\mathbb{E}\left(\text{Conf}_{ORIG}(x^{(\mathbb{T})})\right)$ is very small (less than 0.1%), which implies that the confidence estimation on complete inputs are very close.

B.6. Confidence Estimation when Just Penalizing the Confidence Difference



(a) Modality 1 (b) Modality 2 (c) Complete modalities

Figure 7: Confidence estimation when penalizing the confidence difference (Eq. 3).

Forcing the confidence for $x^{(\mathbb{T})}$ to be smaller than the confidence for $x^{(\mathbb{S})}$ strictly (Eq. 3) will lead to a very small confidence for $x^{(\mathbb{T})}$ and will make the model estimate an extremely small confidence for each modality, which contradicts the fact that the model sometimes can still make correct predictions confidently when one modality is removed. A flexible ranking regularization makes it more suitable for real data.

C. Analysis of the Training Time and Space Complexity

Ideally, CML should be computed over all possible pairs at each model update. However, it is computationally expensive, so we employ an approximation scheme following (Toneva et al., 2018) for reducing the costs. For example, given samples with 4 modalities (a, b, c, d), we need to sample 3 pairs (a/ab, ab/abc, abc/abcd) to approximate CML loss, and indexes are shuffled for different epochs. So if the complexity of the traditional model is o(n), the complexity of our method will be o((k-1)n), where k indicates the number of modalities. It should be pointed out that compared models in our experiments are also equipped with sampling (to avoid the influence of sampling), and the complexity of compared methods is also o((k-1)n). We report the training time (seconds) for the same training epochs (Platform: RTX 3090×8, CUDA Version: 11.2). It is observed that the original model and model equipped with CML have the same level of computational complexity.

Table 7: Training time (Platform: RTX 3090 ×8).

Method	CML	TUANDROMD	YaleB	Handwritten	CUB	Animal
Type I	✗	245.3	1574.6	141.5	351.6	1582.7
	✓	297.6	1210.2	191.2	348.5	1641.3
Type II	✗	1447.7	703.3	233.2	565.2	717.8
	✓	1489.1	662.9	210.8	781.7	720.3

D. Algorithms

In addition to the general algorithm shown in the main text, we show the specific algorithms corresponding to different types of algorithms and add more comments for better understanding.

D.1. CML for Imputation-independent Model

Algorithm 2 CML for the imputation-independent model

```
Given dataset $\mathcal{D} = \left\{\{x_i^m\}_{m=1}^M, y_i\right\}_{i=1}^N$, classifier $f$, and classification loss function $\mathcal{L}^{\text{CL}}$, Coefficient $\lambda$ of CML, epochs for training the classifier $epoch$
for $e = 1, \ldots, epoch$ do
  $\mathbb{S} \leftarrow \mathbb{M}$
  Make the prediction via input $\mathbb{S}$
  $\mathcal{L}^{\text{CL}} \leftarrow \mathcal{L}^{\text{CL}}(x^{(\mathbb{S})})$
  $\mathcal{L}^{\text{CML}} \leftarrow 0$
  for $m = M - 1, \ldots, 1$ do
    Randomly erase a modality of $\mathbb{S}$ and set it as $\mathbb{T}$
    Make the prediction via input $\mathbb{T}$
    $\mathcal{L}^{\text{CL}} \leftarrow \mathcal{L}^{\text{CL}} + \mathcal{L}^{\text{CL}}(x^{(\mathbb{T})})$
    $\mathcal{L}^{\text{CML}} \leftarrow \mathcal{L}^{\text{CML}} + \max\left(0, \text{Conf}(x^{(\mathbb{T})}) - \text{Conf}(x^{(\mathbb{S})})\right)$
  end for
  $\mathcal{L} = \frac{1}{M}\mathcal{L}^{\text{CL}} + \lambda\mathcal{L}^{\text{CML}}$
  Update the parameters of the classification model with $\mathcal{L}$
end for
return the classifier $f_{\text{CL}}$
```

D.2. CML for Imputation-dependent Model

For imputation-dependent method, we use MIWAE to train the reconstruction model first, then we use the reconstructed modalities to train the classifier.

For reconstruction-based method, the missing modalities need to be reconstructed first, so the process can be divided into two stages.

E. Discussion

E.1. Class-imbalanced

○ Why the CML can still work when the training data is class-imbalanced (e.g., long-tailed)?

CML can improve performance when the data for the training model is class-imbalanced since it increases the confidence of the minority classes. For a trustworthy model, the model should treat the majority and minority classes equally during the test. CML requires the model to make predictions fairly regardless of whether the majority and minority classes of the samples belong. On the contrary, the original model tends to predict lower confidence for the minority classes than the majority classes. And the improvements on the class-imbalanced dataset Animal (data distribution is shown in Fig. 8) validate the effectiveness.

Animal is a class-imbalanced real-world dataset, the improvement shows CML can also deal with applications that suffer from class-imbalanced. The original model tends to predict lower confidence for the minority classes than the majority classes, which is unfair to minority classes. CML requires the model to make predictions fairly regardless of whether the majority and minority classes of the samples belong.

E.2. Pair-wise Sampling

The exact computation of the proposed loss needs to enumerate all modality set pairs (i.e., $\mathbb{T}$ and $\mathbb{S}$), which is typically computational expensive sometimes. Therefore, we introduce a strategy (Moon et al., 2020; Toneva et al., 2018) to approximate this loss by sampling modality set pairs and find this strategy works well in practice. If the complexity of the traditional model is o(n), the complexity of our method will be o((k-1)n), where k indicates the number of modalities.

Algorithm 3 CML for the imputation-dependent model

```
Given dataset $\mathcal{D} = \left\{\{x_i^m\}_{m=1}^M, y_i\right\}_{i=1}^N$, reconstruction network $f_{\text{re}}$ and classifier $f_{\text{CL}}$, reconstruction loss function $\mathcal{L}^{\text{re}}$,
Coefficient $\lambda$ of CML, epochs for training the reconstruction net $epoch_{\text{re}}$ and classifier $epoch_{\text{CL}}$
for $e_1 = 1, \dots, epoch_{\text{re}}$ do
    Reconstruct the modalities via reconstruction model
    Compute the reconstruction loss by $\mathcal{L}^{\text{re}}$
    Update the parameters of the reconstruction model
end for
for $e_2 = 1, \dots, epoch_{\text{CL}}$ do
    $\mathbb{S} \leftarrow \mathbb{M}$
    $\mathcal{L}^{\text{CE}} \leftarrow \mathcal{L}^{\text{CE}}(x^{(\mathbb{S})})$
    $\mathcal{L}^{\text{CML}} \leftarrow 0$
    for $m = M-1, \dots, 1$ do
        Randomly erase a modality of $\mathbb{S}$ and set it as $\mathbb{T}$
        Reconstruct the erased modalities via reconstruction model and add them to $x^{(\mathbb{T})}$
        Compute the classification loss $\mathcal{L}^{\text{CE}}(x^{(\mathbb{T})})$ with Cross-Entropy loss function
        $\mathcal{L}^{\text{CE}} \leftarrow \mathcal{L}^{\text{CE}} + \mathcal{L}^{\text{CE}}(x^{(\mathbb{T})})$
        $\mathcal{L}^{\text{CML}} \leftarrow \mathcal{L}^{\text{CML}} + \max\left(0, \text{Conf}(x^{(\mathbb{T})}) - \text{Conf}(x^{(\mathbb{S})})\right)$
    end for
    $\mathcal{L} = \frac{1}{M}\mathcal{L}^{\text{CE}} + \lambda\mathcal{L}^{\text{CML}}$
    Update the parameters of the classification model with $\mathcal{L}$
end for
return the reconstruction model $f_{\text{re}}$ and classifier $f_{\text{CL}}$
```

F. CML being Deployed in Advanced Multimodal Models

MMTM is a state-of-the-art method in multimodal classification which is selected as a representative method by (Wu et al., 2022) and originally proposed by (Joze et al., 2020). NYU Depth V2 and SUN RGB-D are two widely used multimodal datasets for RGB-D scene recognition. ○ **NYUD2**: Following previous work (Georghiades et al., 2002), we use a reorganized version of this dataset, which contains 1449 samples, 10 scene classes. ○ **SUN RGB-D** (Perkins & Theiler, 2003): This is a standard database of RGB-D scene recognition. Similar to previous work (Georghiades et al., 2002), we also use a subset of this dataset which contains the 19 major scene categories and 9504 samples in total. Following the author's implementation, We employ pre-trained ResNet-18 as the backbone network for MMTM. The input images are fed into depth and visual block first. Then the rgb and depth features are fused by MMTM before the final prediction. We add CML regularization to the softmax output before and after MMTM fusion process. In our experiment, the squeeze ratio of MMTM Module is set to 16. The dimensionalities of rgb and depth feature are both 512.

G. Related Work Details

Uncertainty estimation provides a way for trustworthy prediction (Abdar et al., 2021). Uncertainty can be used as an indicator of whether the predictions given by models are prone to be wrong. Many uncertainty-based models have been proposed in the past decades, such as Bayesian neural networks (Neal, 2012; MacKay, 1992; Denker & LeCun, 1990; Kendall & Gal, 2017), Dropout (Molchanov et al., 2017), and Deep ensembles (Lakshminarayanan et al., 2017; Havasi et al., 2020). Built upon RBF networks, DUQ (van Amersfoort et al., 2020) is able to identify the out-of-distribution samples, which uses distance to represent the prediction uncertainty. Prediction confidence is always referred to in classification models, which expects the predicted class probability to be consistent with the empirical accuracy. Models are frequently overconfident because softmax probabilities are computed with the fast-growing exponential function (Hendrycks & Gimpel, 2017), so many methods focus on smoothing the prediction probabilities distribution, such as Label smoothing (Müller et al., 2019). The recent approach employs the focal loss to calibrate the deep neural networks (Mukhoti et al., 2020). A recent work (Corbière et al., 2019) introduces True Class Probability (TCP) to ensure the low confidence for the failure predictions. Temperature scaling (TS) (Guo et al., 2017) is a well-known post-hoc confidence calibration method, which aims to re-scale the output probability by manipulating the softmax inputs, i.e., the logits.

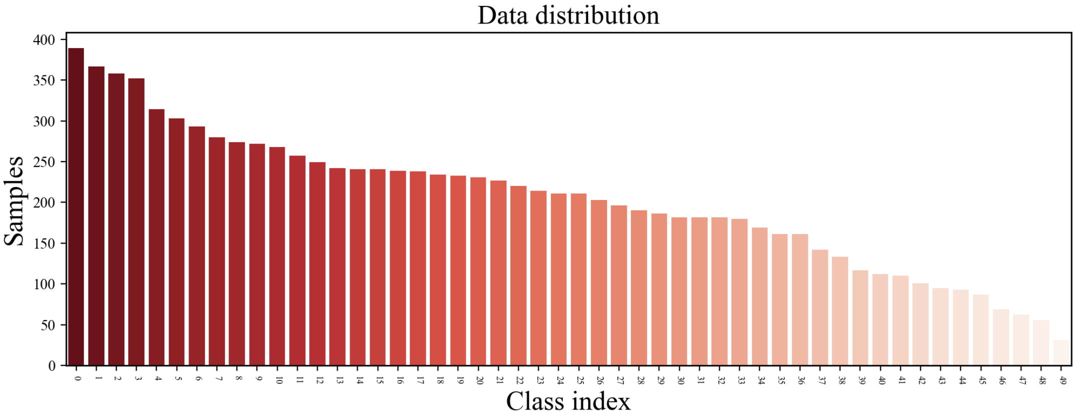


Figure 8: Illustration of data distribution of Animal dataset (the number of samples for every classes).

Recently, there have been a wide range of research interests in handling missing modalities for multimodal learning, including imputation-independent methods (Zhang et al., 2019) and imputation-dependent methods (Mattei & Frellsen, 2019; Wu & Goodman, 2018). Imputation-independent methods have no need to reconstruct the missing modalities and make classification via an uniform representation. For imputation-dependent methods (based on reconstruction), the strategy model can be split into two stages, reconstructing the missing modalities and making classification according to the reconstructed modalities. CPM-Nets (Zhang et al., 2019) is an advanced method which can guarantee the performance by fully exploiting all samples and all modalities to produce structured representation for interpretability, and the method has been extended and deployed into medical domain (Lee & van der Schaar, 2021). MIWAE (Mattei & Frellsen, 2019) is a typical reconstruction model in multimodal classification, whose objective is a lower bound of the likelihood of the observed data that can be tight in the limit of very large computational power.

H. Refinement and modification following peer review

H.1. Underlying reason of why the confidence violates the condition

(1) The most likely reason is the "greedy" nature of multimodal learning. Prior research (Han et al., 2021) has acknowledged that multimodal learning models often exhibit over-reliance on certain modalities while under-training on others, resulting in over-confidence on one input modality and an increase in confidence (statistically) when other modalities are removed.

(2) To verify this hypothesis, we assessed whether the degree of "greediness" (as defined in (Han et al., 2021)) and VRR are positively correlated using the Pearson correlation coefficient. We trained models with various seeds and consistently observed confidence violations in "greedy" models, as shown in the table below. Pearson correlation coefficient between VRR and Greedy (Wu et al., 2022) on SOTA method.

(3) This finding supports the notion that the proposed regularization can enhance multimodal models by mitigating their inherent greediness. Future research will explore the theoretical link between VRR and Greedy.

H.2. Differences from traditional calibration metrics

The proposed metric is distinct from external metrics that utilize class labels, as it is the first internal metric designed to assess calibration. The differences between external metrics and internal metrics can be analogous to clustering metrics.

(1) The proposed metric is an internal metric, while ECE and Brier score are external metrics.

(2) External metrics using class labels evaluate whether the model's confidence and accuracy are aligned from a global classification perspective. The proposed internal metric, however, is labels-free and assesses whether a model inherently meets certain criteria.

(3) We anticipate that additional internal metrics will be introduced in the future, analogous to the clustering field, and this work will benefit the community.

Table 8: Accuracy performance comparison of MMTM when some of the modalities is corrupted with color jitter (i.e., randomly change the brightness, contrast, saturation and hue of an image with jitter factor ϵ.).

Dataset	Noise on	CML	$\epsilon = \mathbf{0.1}$	$\epsilon = \mathbf{0.2}$	$\epsilon = \mathbf{0.3}$	$\epsilon = \mathbf{0.5}$
NYUD-2	{1}	✗	65.72 ± 0.70	64.13 ± 1.78	63.79 ± 1.79	60.89 ± 1.21
		✓	66.64 ± 1.22	65.41 ± 0.65	64.31 ± 0.92	62.26 ± 1.77
		Improve	△ 0.92	△ 1.28	△ 0.52	△ 1.37
	{2}	✗	61.34 ± 0.98	57.98 ± 0.81	53.98 ± 2.28	52.26 ± 3.23
		✓	62.63 ± 0.60	57.89 ± 1.56	54.80 ± 2.90	52.57 ± 3.38
		Improve	△ 1.29	▽ 0.09	△ 0.82	△ 0.31
	{1, 2}	✗	60.43 ± 0.82	55.17 ± 0.85	51.01 ± 2.64	41.52 ± 4.01
		✓	61.87 ± 0.93	56.24 ± 2.22	51.53 ± 1.91	41.99 ± 3.37
		Improve	△ 1.44	△ 1.07	△ 0.52	△ 0.47
SUN-RGBD	{1}	✗	60.72 ± 0.58	58.98 ± 0.72	57.40 ± 0.75	55.68 ± 0.95
		✓	61.50 ± 0.59	59.95 ± 0.17	57.97 ± 0.30	57.21 ± 0.32
		Improve	△ 0.78	△ 0.97	△ 0.57	△ 1.53
	{2}	✗	60.11 ± 0.24	58.57 ± 0.60	57.46 ± 0.69	55.25 ± 1.05
		✓	59.90 ± 0.49	58.44 ± 0.75	57.25 ± 0.56	55.34 ± 0.87
		Improve	▽ 0.21	▽ 0.13	▽ 0.21	–
	{1, 2}	✗	58.67 ± 0.42	54.77 ± 0.44	51.66 ± 0.64	45.68 ± 1.35
		✓	58.95 ± 0.20	54.73 ± 0.71	51.36 ± 0.66	45.99 ± 1.24
		Improve	△ 0.28	–	▽ 0.30	△ 0.31

H.3. Connection to unbalanced multimodal problem

(1) The proposed method can address the problem of relying on partial modalities, as demonstrated in Table 4 and Table 5 in Appendix.

(2) The model becomes more robust when one of the modalities is corrupted, which can be considered as unbalanced multimodal problem.

(3) We evaluate the relationship between the VRR and Greedy (defined in (Wu et al., 2022) which indicates the degree of over-relying on a certain modality) by calculating the Pearson correlation coefficient according to different seeds. Pearson correlation coefficients between VRR and Greedy on SOTA method (i.e., MMTM) are 0.940 and 0.915 on NYUD2 and SUN-RGBD dataset respectively. According to empirical results, confidence violation always occurs with "greedy".

H.4. Analysis of loss function sampling approach

(1) In practice, enumerating all pairs would involve permutation and combination, making it computationally expensive (detailed complexity analyses can be found in Appendix E.2).

(2) Hence, we use a sampling strategy to approximate the loss function, as demonstrated in Appendix A. The sampling approach has been widely used in various methods that encounter the same problem (Toneva et al., 2018; Moon et al., 2020), and has shown good approximation ability and stability.

(3) In our experiments, we introduce this sampling approach since it is widely used.

H.5. Analysis of hyper parameters

(1) We choose the value of that achieves the best performance on the validation set 1, 5, 10, . . . , 100.

(2) Moreover, as demonstrated in the ablation study (Fig. 4), the proposed regularization is not sensitive to the hyperparameter.

Table 9: Accuracy performance comparison of MMTM when some of the modalities is corrupted with gaussian noise (i.e., zero mean with varying variance ϵ).

Dataset	Noise on	CML	$\epsilon = \mathbf{0.1}$	$\epsilon = \mathbf{0.2}$	$\epsilon = \mathbf{0.3}$	$\epsilon = \mathbf{0.5}$
NYUD-2	$\{1\}$	✗	64.77 ± 1.76	63.03 ± 1.92	61.50 ± 2.83	58.81 ± 4.05
		✓	65.26 ± 1.92	63.98 ± 1.60	62.94 ± 1.97	59.88 ± 3.03
		Improve	△ 1.49	△ 0.95	△ 1.44	△ 1.07
	$\{2\}$	✗	65.41 ± 1.27	62.17 ± 1.76	59.08 ± 1.54	55.75 ± 2.75
		✓	66.12 ± 1.10	62.75 ± 1.26	59.79 ± 2.23	55.90 ± 3.38
		Improve	△ 1.29	△ 0.58	△ 0.71	△ 0.15
	$\{1, 2\}$	✗	61.87 ± 0.82	55.60 ± 2.61	48.62 ± 4.32	37.68 ± 4.94
		✓	63.12 ± 1.49	57.31 ± 1.58	49.51 ± 2.75	37.98 ± 5.21
		Improve	△ 1.25	△ 1.71	△ 0.89	△ 0.30
SUN-RGBD	$\{1\}$	✗	60.69 ± 0.65	58.78 ± 0.95	56.84 ± 1.13	53.14 ± 1.32
		✓	61.00 ± 0.32	59.31 ± 0.83	57.47 ± 0.62	54.77 ± 1.00
		Improve	△ 0.31	△ 0.53	△ 0.63	△ 1.63
	$\{2\}$	✗	60.93 ± 0.58	59.25 ± 0.71	57.55 ± 1.08	54.81 ± 1.66
		✓	61.25 ± 0.59	59.19 ± 0.68	57.50 ± 1.27	54.34 ± 1.93
		Improve	△ 0.32	–	–	▽ 0.47
	$\{1, 2\}$	✗	59.16 ± 0.88	53.56 ± 1.51	47.22 ± 2.12	35.90 ± 2.38
		✓	59.59 ± 1.09	54.14 ± 0.58	47.38 ± 1.47	36.30 ± 2.39
		Improve	△ 0.43	△ 0.58	△ 0.16	△ 0.40

Table 10: VRR (%) of test samples (a lower value indicates a better confidence estimation). "✗" indicates the model is not equipped with the proposed regularization ($\lambda = 0$).

Method	CML	NYUD-2	SUN-RGBD
Type III	✗	58.09 ± 4.46	57.09 ± 1.50
	✓	46.99 ± 2.89	52.56 ± 3.49
	Improve	△ 11.10	△ 4.53

Table 11: Accuracy under different λ

Model	Dataset	$\lambda = \mathbf{10.0}$	$\lambda = \mathbf{20.0}$	$\lambda = \mathbf{30.0}$	$\lambda = \mathbf{50.0}$	$\lambda = \mathbf{100.0}$
CPM	Animal	81.83 ± 2.58	82.56 ± 1.69	82.73 ± 1.64	82.57 ± 1.78	82.30 ± 2.08
	CUB	86.67 ± 4.68	88.33 ± 4.05	86.33 ± 5.49	87.17 ± 3.05	87.17 ± 3.44
MIWAE	Animal	86.91 ± 0.39	87.40 ± 0.20	87.41 ± 0.38	87.24 ± 0.30	87.32 ± 0.12
	CUB	93.83 ± 1.63	93.50 ± 1.78	93.67 ± 2.02	97.50 ± 1.33	93.16 ± 2.07

Promising performance can be achieved with a mild regularization strength, indicating that the proposed regularization is not sensitive to hyperparameters and can be easily deployed in a wide range of multimodal models using CML.