CALIBRATING MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

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 methods suffer from unreliable predictive confidence that tends 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 classification, 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

Multiple modalities 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 (e.g., instance classification). Numerous approaches have been proposed to build multimodal classification 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 classification methods remains largely unexplored. One key aspect of the reliability is to build a high-quality uncertainty estimator (Neal, 2012; MacKay, 1992), which can quantitatively characterize the probability that predictions will be wrong. 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 classification, in addition to exact overall prediction confidence, the relationship between the modalities should also be taken into concerns. Intuitively, the confidence of an ideal multimodal classifier should not increase when one modality is removed. 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, in practice, the confidence estimation obtained by ordinary training methods tends to be overconfident to partial modalities, which violates the principle and leads to some counter-intuitive phenomenons. We perform extensive empirical studies and observe that when a modality is removed, the overall confidence can even increase. This observation contradicts the usual assumption of multimodal classification since modalities are assumed to be predictive of the target for most multimodal classification tasks (Wu et al., 2022). 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 taking all modalities into account fairly.

A natural idea to address the above issue is to employ recent 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 uncertainty estimation than the

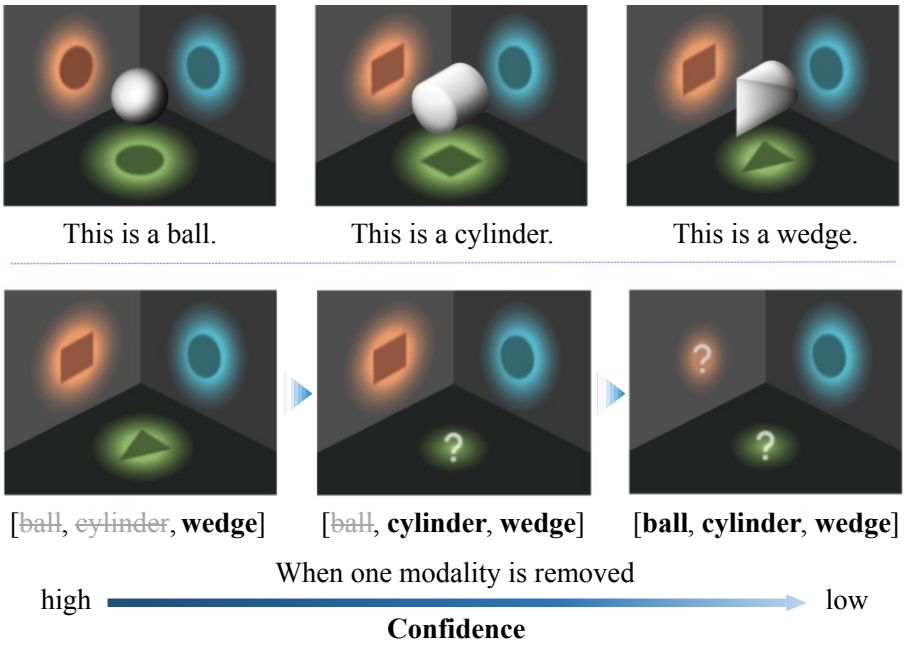


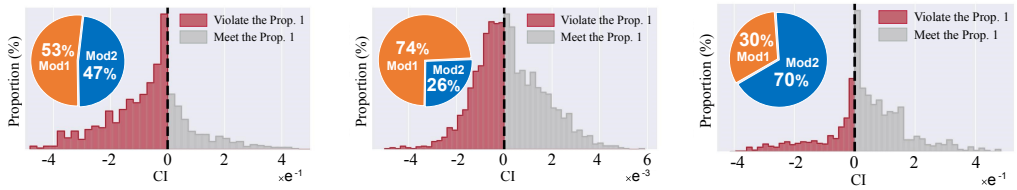
Figure 1: Motivation of calibrating multimodal learning. The confidence of an ideal multimodal classifier should decrease (at least not increase) when one modality is removed.

traditional training/inference manner. However, these approaches do not explicitly consider the relationship between different modalities (i.e., they can only calibrate the fused confidence but cannot adjust the relationship between different modalities during training) and thus still fail to achieve satisfactory performance in the multimodal classification setting. Therefore, we claim that high-quality uncertainty estimation in various multimodal classification tasks needs to explicitly treat all modalities fairly. To this end, 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 perform extensive empirical studies to show that most existing multimodal classification paradigms tend to over-rely on partial modalities (different samples over-rely on different modalities rather than all samples over-rely on the same modalities), which implies they fail to achieve trustworthy uncertainty 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 classification methods. We 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). 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



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

Figure 2: Current methods (MMTM (Wu et al., 2022), CPM-Nets (Zhang et al., 2019), and MIWAE (Mattei & Frellsen, 2019)) violate the Proposition 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 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").

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** is always referred to in classification models, which expects the predicted class probability to be consistent with the empirical accuracy. 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 train a machine learning model. Recently, there have been a wide range of research interests in handling missing modalities for multimodal learning, including imputation-independent (Type I) methods (Zhang et al., 2019) and imputation-dependent (Type II) 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. Besides the methods for incomplete multimodal learning, recent multimodal methods (Type III) Joze et al. (2020); Wu et al. (2022) achieve SOTA performance in multimodal video classification. We evaluate the performance of CPM-Nets (Zhang et al., 2019), MIWAE (Mattei & Frellsen, 2019), and MMTM (Joze et al., 2020; Wu et al., 2022) in experiments due to their representativeness in different types of multimodal learning.

3 METHOD

In this section, we first introduce some basic notations in Section 3.1. Then, we show the basic assumption of our method and its empirical motivations in Section 3.2 and evaluate the confidence estimation performance of current multimodal methods in Section 3.3. At the end, we propose a simple yet effective regularization technique 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 input feature of the m-th modality of the i-th sample, and $y_i \in \{1, \cdots, K\}$ is the corresponding class label. To distinguish whether the input is a unique modality or a set of modalities, we use x^m to represent the m-th modality, and use the $\mathbf{x}^{(\mathbb{S})}$ to represent multimodal input set, where $\mathbb{S}$ is a set of modalities' indexes (e.g., if we have $\mathbb{S} = \{1, 2\}$, then $\mathbf{x}^{(\mathbb{S})}$ indicates a feature set consisting of x^1 and x^2, and $\mathbf{x}^{(\mathbb{M})} = \{\mathbf{x}^1, \cdots, \mathbf{x}^M\}$ indicates the complete M modalities). The goal is to learn a function: $f(\mathbf{x}^{(\mathbb{M})}) \rightarrow z$, 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 $\mathrm{P}(y \mid \mathbf{w}, \mathbf{x}^{(\mathbb{M})}) = \{\hat{p_k}\}_1^K$. The predicted class label is: $\hat{y} = \arg\max_y \mathrm{P}(y \mid \mathbf{w}, \mathbf{x}^{(\mathbb{M})})$.

3.2 BASIC ASSUMPTION

In the real-world applications, the quality of test samples is unstable (e.g., some modalities may be corrupted), so the quality (i.e., available modalities) of the multimodal input should be reflected in some quantitative manner (i.e., predictive confidence) when multimodal methods are deployed in the real-world applications. However, it is difficult to exactly define the "quality" of each sample, and we cannot define the exact functional relationship between the quality and confidence since the confidence estimation is unique for different models even for a same sample. To address this issue, we approximate this relationship with a ranking-based form as follow:

Proposition 1. *Given two versions of a unique sample $\mathbf{x}^{(\mathbb{M})}$, i.e., $\mathbf{x}^{(\mathbb{T})}$ and $\mathbf{x}^{(\mathbb{S})}$, if we can assure $\mathbb{T} \subset \mathbb{S}$, then, for a trustworthy multimodal classifier $f(\cdot)$, it should hold* $\mathrm{Conf}(f(\mathbf{x}^{(\mathbb{T})})) \leq \mathrm{Conf}(f(\mathbf{x}^{(\mathbb{S})}))$.

Proposition 1 indicates that the predictive confidence shouldn't increase when one modality is removed. We further define the prediction **C**onfidence **I**ncrement (CI) with informativeness increment for a unique sample as:

$$\mathrm{CI}(\mathbf{x}^{(\mathbb{S})}, \mathbf{x}^{(\mathbb{T})}) = \mathrm{Conf}(f(\mathbf{x}^{(\mathbb{S})})) - \mathrm{Conf}(f(\mathbf{x}^{(\mathbb{T})})) \quad \text{subject to: } \mathbb{T} \subset \mathbb{S}, \tag{1}$$

where $\mathbb{T}$ and $\mathbb{S}$ are sets of modalities' indexes. Specially, a negative value indicates a poor confidence estimation performance that the predictive confidence increases when one modality is removed. To quantify the extent that a learned model violates Proposition 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:

$$\mathrm{VRR} = \mathbb{E}\left[\mathbb{1}\left(\mathrm{CI}(\mathbf{x}^{(\mathbb{S})}, \mathbf{x}^{(\mathbb{T})}) < 0\right)\right] \quad \text{subject to: } \mathbb{T} \subset \mathbb{S}. \tag{2}$$

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 incomplete multimodal, we also evaluate the MMTM (Wu et al., 2022), which is a SOTA multimodal classification method. As shown in Tab. 1, the VRR scores of previous methods are quite high which indicates the prediction confidence on many samples will violate Proposition 1. And the visualization is shown in Fig. 2, where the red color indicates the proportion of test samples whose predictive confidence given 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 over-rely on during the fusion). As shown in Fig. 2, however, we find that different samples over-rely on different modalities rather than all samples over-rely on the same modality. This indicates that the problem can't 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 term of simple-wise, 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, which implies that the confidence and its quality are not matched. For this issue, they cannot be deployed into risk-sensitive applications such as medical analysis. As a comparison, our method can significantly decrease VRR score (see more details in Tab. 1) for a more trustworthy confidence estimation.

3.4 CML REGULARIZATION

As shown in Section 3.3, the current multimodal methods usually increase the prediction confidence when one modality is removed, which hinders the model performance inherently. To address this issue, a naive strategy is to penalize the confidence difference between the $\mathbf{x}^{(\mathbb{T})}$ and $\mathbf{x}^{(\mathbb{S})}$:

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

However, models sometimes can still make an accurate prediction confidently when one modality is removed in practice. Eq. 3 forces the models to predict small confidence strictly when one modality is removed, which pushes the model to estimate an extremely small confidence for each modality (the illustration please refer to Appendix C.6). 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 $\mathbb{T} \subset \mathbb{S}$, the regularization can be written as:

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

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

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

The exact computation of above loss needs to enumerate all modality set pairs (i.e., $\mathbb{T}$ and $\mathbb{S}$), which is typically computational expensive sometimes. Therefore, we propose to approximate this loss by sampling modality set pairs and find this strategy works well in practice. Specifically, we conduct sampling as same as that in computing VRR (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, the objective function is induced as:

$$\mathcal{L}_i = \mathcal{L}_i^{\mathrm{cl}} + \lambda \mathcal{L}_i^{\mathrm{CML}}, \tag{6}$$

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

Algorithm 1: The training pseudocode for deploying CML regularization

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

3.5 DISCUSSION AND ANALYSES

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

Lemma 3.1. *Suppose we have two versions of a unique sample* $\mathbf{x}^{(\mathbb{M})}$*, i.e.,* $\mathbf{x}^{(\mathbb{T})}$ *and* $\mathbf{x}^{(\mathbb{S})}$*, if we can assure* $\mathbb{T} \subset \mathbb{S}$*, then, for any class label* y*, we have* $I(y, \mathbf{x}^{(\mathbb{T})}) \leq I(y, \mathbf{x}^{(\mathbb{S})})$.

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

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

Lemma 3.2. *Suppose the CML regularization can achieve a lower* VRR*, i.e.,* $\mathrm{VRR}_{CML} < \mathrm{VRR}_{Ori}$*, then for the samples that meet* $\mathbb{E}\left(\mathrm{Conf}_{CML}(\mathbf{x}^{(\mathbb{S})})\right) = \mathbb{E}\left(\mathrm{Conf}_{Ori}(\mathbf{x}^{(\mathbb{S})})\right)$*, we have* $\mathbb{E}\left(\mathrm{Conf}_{CML}(\mathbf{x}^{(\mathbb{T})})\right) \leq \mathbb{E}\left(\mathrm{Conf}_{Ori}(\mathbf{x}^{(\mathbb{T})})\right)$.

Although it is difficult to make $\mathrm{Conf}_{Ori}(\cdot)$ equal to $\mathrm{Conf}_{CML}(\cdot)$ strictly for all samples, as shown in Appendix C.5, we find $\mathrm{Conf}_{CML}(\mathbf{x}^{(\mathbb{S})})$ and $\mathrm{Conf}_{Ori}(\mathbf{x}^{(\mathbb{S})})$ are very similar for most samples, where $\mathrm{Conf}_{Ori}(\cdot)$ and $\mathrm{Conf}_{CML}(\cdot)$ indicate the confidence estimated by the original model and the model improved by CML regularization respectively.

◦ **Why not just penalize the difference in confidence:** $\mathrm{Conf}(\mathbf{x}^{(\mathbb{T})}) - \mathrm{Conf}(\mathbf{x}^{(\mathbb{S})})$? Forcing the confidence for $\mathbf{x}^{(\mathbb{T})}$ to be smaller than the confidence for $\mathbf{x}^{(\mathbb{S})}$ strictly will lead to a very small confidence for $\mathbf{x}^{(\mathbb{T})}$ and will make the model estimate an extremely small confidence for each modality, and the experiments are shown in Appendix C.6. What's more, 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.

4 EXPERIMENTS

4.1 SETUP

We deploy the proposed regularization method to different types of multimodal classifiers including the imputation-independent method (Type I), the imputation-dependent method (Type II), and the recent SOTA 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 typical imputation-dependent algorithm. The above two methods are typical in incomplete multimodal learning. In addition to incomplete multimodal learning methods, we also deploy the regularization to the current multimodal method (Wu et al., 2022), which is named MMTM. For MMTM, we approximate removing one modality by corrupting its features (e.g., adding strong noise) due to the model can't make a prediction when one modality is absolutely removed.

Datasets: We evaluate the proposed method on diverse datasets, including YaleB (Georghiades et al., 2002), Handwritten (Perkins & Theiler, 2003), CUB (Wah et al., 2011), Animal (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014), TUANDROMD (Borah et al., 2020), NYUD2 (Qi et al., 2017), and SUNRGBD (Song et al., 2015).

Experiment setting: For a fair comparison, the only difference between whether the model is equipped with CML regularization or not is whether the coefficient λ is set to 0. Please refer to Appendix C.2 for more detailed settings.

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 methods on five datasets and evaluate their trustworthiness in terms of VRR (defined in the 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 C.2.

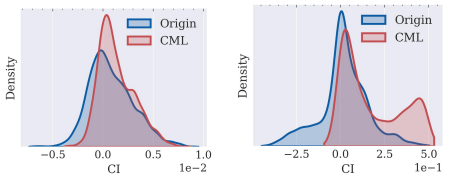
Table 1: VRR (%) of test samples (a lower value indicates a better confidence estimation. Type III is shown in Appendix). "✗" indicates the model is not equipped with the proposed regularization ($\lambda = 0$). Performance on Type III please refer to Tab. 8 (Appendix F).

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

4.3 Results

4.3.1 Confidence estimation

We evaluate the confidence estimation of current multimodal classification models from a ranking perspective and find that for a large number of samples the confidence will increase when one modality is removed, while the confidence estimation of classification models equipped with the proposed CML regularization is significantly improved. We intuitively demonstrate the confidence changing in Fig. 3, 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 doesn't 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.



(a) CUB (b) Tuandromd

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

We also report the quantitative results on five datasets. It is observed that the confidence estimation of each model is obviously improved with the proposed CML regularization.

Table 2: Accuracy performance comparison for whether the model is equipped with the CML regularization term (i.e., whether λ is set to 0).

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	△ 0.86	△ 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	△ 0.05	△ 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
	SUNRGBD	✗	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

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 counterpart (i.e., the original classification models) validating the rationality of CML principle. Limited by space, results on more datasets are shown in Appendix C.4.

We also 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 C.3. We can find that the models equipped with CML regularization are more robust to noise, especially when the noise is much heavier.

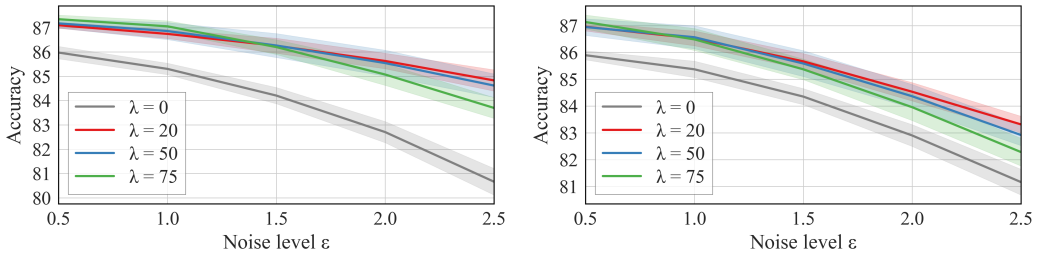
4.3.3 PERFORMANCE UNDER DIFFERENT STRENGTHS OF CML REGULARIZATION

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, and it is clear that the model with CML regularization is more robust to the potential noise.

We show the results in Fig. 4. From Fig. 4, we can find that CML regularization can promote the accuracy on the noised data. The main reason is that the CML regularization enforces the reasonable

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



(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.

confidence estimation. Moreover, according to Fig. 4, the proposed regularization is not sensitive to the hyperparameter λ, where quite a promising performance could be excepted with a mild regularization strength.

5 Conclusion

Through extensive empirical studies, we observe that the confidence estimations of current multimodal learning algorithms are typically unreliable, which tend to rely on some partial modalities. This further leads the learned model to being non-robust against the modality corruption. To be specific, model tends to be overconfident to some modalities, and ignores the evidences from other modalities even those may be useful to make decision. To solve this problem, we introduce a novel regularization technique to calibrate the confidence estimation, which forces model to learn a calibrated predictive distribution. This technique can be naturally applied to most existing multimodal learning methods without modifying their original training process and model structures. We perform comprehensive experiments to demonstrate our method's superiority in classification performance, confidence calibration and model robustness.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

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

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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.

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

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

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

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *NeurIPS*, 2019.

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

Andrew Foong, David Burt, Yingzhen Li, and Richard Turner. On the expressiveness of approximate inference in bayesian neural networks. *NeurIPS*, 33:15897–15908, 2020.

A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. 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.

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

Danijar Hafner, Dustin Tran, Timothy P. Lillicrap, Alex Irpan, and James Davidson. Noise contrastive priors for functional uncertainty. In *UAI*, 2019.

Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew M Dai, and Dustin Tran. Training independent subnetworks for robust prediction. *arXiv preprint arXiv:2010.06610*, 2020.

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

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13289–13299, 2020.

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

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

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

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, volume 25, 2012.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

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

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

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

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

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

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In *NeurIPS*, 2020.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *NeurIPS*, 2019.

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

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

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

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

Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for rgbd semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5199–5208, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. 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.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *ICML*, 2020.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Weiyao Wang, Du Tran, and Matt Feiszli. 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.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, 2019.

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

Nan Wu, Stanisław Jastrzębski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, 2022.

Changqing Zhang, Zongbo Han, yajie cui, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Cpm-nets: Cross partial multi-view networks. In *NeurIPS*, volume 32, 2019.

CONTENTS

A HOW TO MAKE RANKING PAIRS

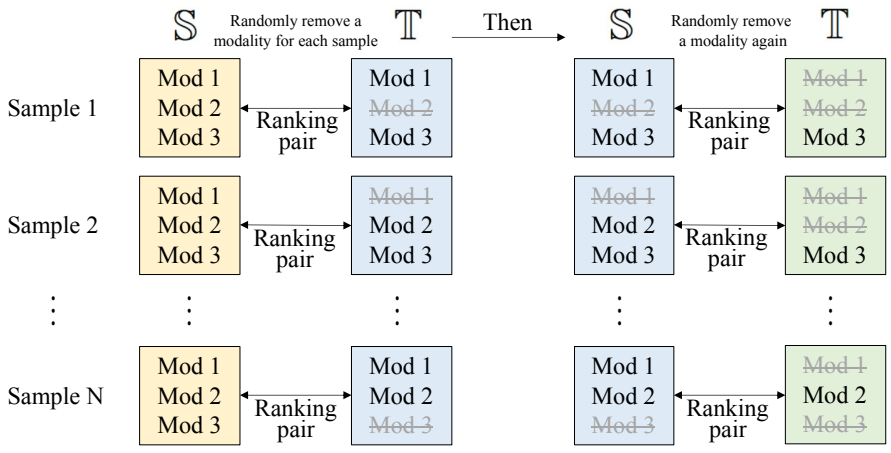


Figure 5: Illustration of generating $\mathbb{S}$ and $\mathbb{T}$.

To compute this score in practice, 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 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, 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 4: 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

C Experiments details

C.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 (2014)). ∘ **TUANDROMD** Borah et al. (2020): The dataset contains $4,465$ instances, 2 classes and 2 types of modalities.

C.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

Table 5: 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
		✓	$\mathbf{98.46 \pm 1.09}$	$\mathbf{98.20 \pm 1.31}$	$\mathbf{96.15 \pm 1.88}$	$\mathbf{94.62 \pm 1.88}$	$\mathbf{93.59 \pm 1.30}$
	{2}	✗	95.13 ± 0.72	94.10 ± 1.31	92.57 ± 0.73	92.05 ± 1.45	91.54 ± 1.66
		✓	$\mathbf{96.92 \pm 1.26}$	$\mathbf{95.90 \pm 2.02}$	$\mathbf{94.61 \pm 2.88}$	$\mathbf{93.33 \pm 2.54}$	$\mathbf{93.08 \pm 3.14}$
	{3}	✗	94.87 ± 0.96	94.87 ± 0.96	94.10 ± 0.96	92.82 ± 1.81	92.05 ± 1.31
		✓	$\mathbf{96.92 \pm 1.88}$	$\mathbf{97.18 \pm 1.92}$	$\mathbf{96.15 \pm 1.88}$	$\mathbf{94.87 \pm 2.54}$	$\mathbf{94.36 \pm 2.02}$
	{1, 2}	✗	96.67 ± 2.61	95.13 ± 3.46	91.28 ± 2.83	88.72 ± 3.10	86.41 ± 3.10
		✓	$\mathbf{97.69 \pm 0.63}$	$\mathbf{95.39 \pm 2.26}$	$\mathbf{92.56 \pm 2.02}$	$\mathbf{89.72 \pm 2.21}$	$\mathbf{86.66 \pm 1.81}$
	{1, 3}	✗	97.43 ± 0.96	97.69 ± 1.66	97.43 ± 1.81	97.18 ± 2.20	96.15 ± 2.26
		✓	$\mathbf{98.46 \pm 1.09}$	$\mathbf{98.46 \pm 1.26}$	$\mathbf{98.46 \pm 1.66}$	$\mathbf{96.92 \pm 1.88}$	$\mathbf{96.67 \pm 2.20}$
	{2, 3}	✗	94.62 ± 1.08	93.85 ± 1.25	90.26 ± 2.54	87.95 ± 2.83	86.67 ± 2.38
		✓	$\mathbf{96.41 \pm 1.81}$	$\mathbf{95.64 \pm 1.92}$	$\mathbf{93.84 \pm 3.32}$	$\mathbf{91.28 \pm 3.10}$	$\mathbf{89.49 \pm 3.16}$
	{1, 2, 3}	✗	96.15 ± 1.88	96.41 ± 3.16	93.85 ± 4.40	87.69 ± 8.21	84.10 ± 10.32
		✓	$\mathbf{97.43 \pm 1.81}$	$\mathbf{97.43 \pm 1.92}$	$\mathbf{93.85 \pm 4.40}$	$\mathbf{87.69 \pm 7.61}$	$\mathbf{82.56 \pm 9.26}$
Hand-written	{1}	✗	97.18 ± 1.92	95.38 ± 1.25	93.34 ± 1.31	92.57 ± 1.58	91.28 ± 1.31
		✓	$\mathbf{98.46 \pm 1.26}$	$\mathbf{95.90 \pm 1.92}$	$\mathbf{93.85 \pm 1.88}$	$\mathbf{93.08 \pm 1.66}$	$\mathbf{92.31 \pm 0.63}$
	{2}	✗	88.46 ± 1.66	87.18 ± 1.31	86.92 ± 1.09	86.92 ± 1.09	86.92 ± 1.09
		✓	$\mathbf{90.77 \pm 3.33}$	$\mathbf{90.26 \pm 3.57}$	$\mathbf{89.75 \pm 3.85}$	$\mathbf{89.75 \pm 3.84}$	$\mathbf{89.75 \pm 3.84}$
	{3}	✗	85.90 ± 1.92	85.13 ± 1.81	84.87 ± 1.45	84.62 ± 1.66	84.62 ± 1.66
		✓	$\mathbf{88.97 \pm 2.54}$	$\mathbf{88.21 \pm 2.61}$	$\mathbf{87.69 \pm 2.74}$	$\mathbf{87.69 \pm 3.32}$	$\mathbf{87.44 \pm 3.10}$
	{1, 2}	✗	88.97 ± 3.68	83.08 ± 3.50	78.97 ± 1.92	77.69 ± 2.74	75.90 ± 3.57
		✓	$\mathbf{88.97 \pm 4.04}$	$\mathbf{83.59 \pm 2.97}$	$\mathbf{80.51 \pm 3.46}$	$\mathbf{77.18 \pm 4.28}$	$\mathbf{74.10 \pm 3.84}$
	{1, 3}	✗	91.54 ± 1.09	91.28 ± 3.16	88.97 ± 5.41	87.43 ± 5.83	85.64 ± 6.42
		✓	$\mathbf{93.59 \pm 2.38}$	$\mathbf{91.79 \pm 3.68}$	$\mathbf{88.97 \pm 4.04}$	$\mathbf{86.93 \pm 4.99}$	$\mathbf{85.39 \pm 4.91}$
	{2, 3}	✗	63.59 ± 8.00	$\mathbf{59.74 \pm 7.00}$	$\mathbf{57.69 \pm 5.99}$	$\mathbf{56.67 \pm 5.94}$	$\mathbf{55.90 \pm 5.49}$
		✓	$\mathbf{64.36 \pm 7.49}$	58.46 ± 6.37	56.67 ± 6.10	55.64 ± 6.04	54.87 ± 6.29
	{1, 2, 3}	✗	54.87 ± 10.68	$\mathbf{37.95 \pm 6.92}$	$\mathbf{29.48 \pm 4.76}$	$\mathbf{24.36 \pm 4.04}$	$\mathbf{22.31 \pm 4.12}$
		✓	$\mathbf{57.18 \pm 11.41}$	35.64 ± 4.80	26.67 ± 2.54	22.82 ± 2.54	20.77 ± 1.09
CUB	{1}	✗	84.72 ± 3.32	82.22 ± 4.53	79.72 ± 4.43	76.39 ± 6.85	71.17 ± 9.14
		✓	$\mathbf{85.83 \pm 2.72}$	$\mathbf{85.00 \pm 3.50}$	$\mathbf{84.17 \pm 4.08}$	$\mathbf{83.06 \pm 3.99}$	$\mathbf{81.11 \pm 4.37}$
	{2}	✗	84.44 ± 2.75	83.89 ± 3.22	83.61 ± 2.83	83.89 ± 3.49	83.61 ± 3.87
		✓	$\mathbf{85.83 \pm 3.40}$	$\mathbf{85.28 \pm 2.75}$	$\mathbf{85.28 \pm 1.97}$	$\mathbf{85.28 \pm 1.97}$	$\mathbf{85.00 \pm 1.80}$
	{1, 2}	✗	85.00 ± 3.12	82.78 ± 3.98	80.00 ± 4.46	76.67 ± 7.48	72.50 ± 11.14
		✓	$\mathbf{85.83 \pm 2.72}$	$\mathbf{85.84 \pm 3.12}$	$\mathbf{85.83 \pm 4.25}$	$\mathbf{84.44 \pm 4.38}$	$\mathbf{81.39 \pm 6.43}$
Animal	{1}	✗	80.78 ± 2.79	80.96 ± 2.78	80.85 ± 2.80	80.81 ± 2.88	80.68 ± 2.93
		✓	$\mathbf{82.03 \pm 1.91}$	$\mathbf{82.37 \pm 2.09}$	$\mathbf{82.55 \pm 2.24}$	$\mathbf{82.42 \pm 2.22}$	$\mathbf{82.30 \pm 2.40}$
	{2}	✗	80.70 ± 2.45	79.81 ± 3.14	77.34 ± 4.80	72.89 ± 7.46	68.52 ± 9.68
		✓	$\mathbf{82.07 \pm 1.57}$	$\mathbf{81.23 \pm 2.32}$	$\mathbf{78.93 \pm 3.65}$	$\mathbf{75.81 \pm 6.30}$	$\mathbf{72.39 \pm 8.35}$
	{1, 2}	✗	80.87 ± 2.55	79.97 ± 3.12	77.11 ± 5.86	72.23 ± 9.04	65.08 ± 12.75
		✓	$\mathbf{82.14 \pm 1.76}$	$\mathbf{81.95 \pm 2.65}$	$\mathbf{79.63 \pm 5.28}$	$\mathbf{76.63 \pm 7.73}$	$\mathbf{72.46 \pm 11.39}$
TUANDROMD	{1}	✗	84.77 ± 0.55	80.47 ± 0.99	76.53 ± 1.11	72.65 ± 0.76	70.17 ± 0.66
		✓	$\mathbf{86.50 \pm 0.59}$	$\mathbf{82.46 \pm 0.77}$	$\mathbf{78.30 \pm 1.18}$	$\mathbf{74.92 \pm 1.39}$	$\mathbf{72.45 \pm 1.33}$
	{2}	✗	86.56 ± 0.27	85.71 ± 0.48	84.14 ± 0.58	82.35 ± 0.86	80.85 ± 1.05
		✓	$\mathbf{88.87 \pm 0.22}$	$\mathbf{88.74 \pm 0.28}$	$\mathbf{88.58 \pm 0.63}$	$\mathbf{88.15 \pm 0.65}$	$\mathbf{87.93 \pm 0.67}$
	{1, 2}	✗	84.88 ± 1.19	80.72 ± 1.02	76.60 ± 0.75	73.15 ± 1.10	70.35 ± 1.25
		✓	$\mathbf{87.41 \pm 3.40}$	$\mathbf{82.78 \pm 1.14}$	$\mathbf{79.28 \pm 1.00}$	$\mathbf{76.30 \pm 1.11}$	$\mathbf{73.82 \pm 1.35}$

prior work Corbière et al. (2019), we evaluated the performance according to Accuracy (%), NLL (10^{-1}), AURC (10^{-3}), and E-AURC (10^{-3}). For both types above, we set 0 as the default value of τ to ensure the model meets CML strictly.

Table 6: Accuracy performance comparison when some of the modalities is blurred (Type II).

Dataset	Noise Noise on	CML	$\epsilon = 0.5$	$\epsilon = 1.0$	$\epsilon = 1.5$	$\epsilon = 2.0$	$\epsilon = 2.5$
YaleB	{1}	✗	95.90 ± 2.54	94.87 ± 3.22	93.85 ± 2.88	93.59 ± 3.16	93.59 ± 3.16
		✓	$\mathbf{97.43 \pm 1.31}$	$\mathbf{96.15 \pm 2.51}$	$\mathbf{95.13 \pm 2.97}$	$\mathbf{94.36 \pm 2.97}$	$\mathbf{93.85 \pm 3.46}$
	{2}	✗	96.15 ± 2.26	93.33 ± 3.22	91.03 ± 2.62	90.26 ± 2.02	89.23 ± 2.18
		✓	$\mathbf{97.69 \pm 1.26}$	$\mathbf{96.67 \pm 1.58}$	$\mathbf{94.10 \pm 2.20}$	$\mathbf{92.82 \pm 2.83}$	$\mathbf{92.05 \pm 2.02}$
	{3}	✗	98.72 ± 0.36	96.92 ± 1.26	96.15 ± 0.63	96.15 ± 0.63	95.90 ± 0.96
		✓	$\mathbf{98.72 \pm 0.73}$	$\mathbf{97.69 \pm 1.09}$	$\mathbf{97.43 \pm 0.96}$	$\mathbf{97.18 \pm 1.31}$	$\mathbf{96.67 \pm 1.58}$
	{1, 2}	✗	95.64 ± 2.83	91.02 ± 3.46	88.46 ± 4.53	87.18 ± 3.46	85.90 ± 4.09
		✓	$\mathbf{96.66 \pm 1.31}$	$\mathbf{93.59 \pm 2.38}$	$\mathbf{90.51 \pm 2.97}$	$\mathbf{86.67 \pm 3.46}$	$\mathbf{84.62 \pm 3.26}$
	{1, 3}	✗	98.46 ± 0.63	98.46 ± 1.66	97.69 ± 1.66	97.43 ± 1.45	97.18 ± 1.31
		✓	$\mathbf{98.20 \pm 0.73}$	$\mathbf{97.95 \pm 1.92}$	$\mathbf{97.69 \pm 1.66}$	$\mathbf{98.20 \pm 1.58}$	$\mathbf{97.69 \pm 1.66}$
	{2, 3}	✗	97.43 ± 0.36	95.89 ± 0.36	95.38 ± 0.62	94.62 ± 0.62	92.82 ± 0.73
		✓	$\mathbf{98.72 \pm 0.36}$	$\mathbf{97.69 \pm 1.09}$	$\mathbf{96.66 \pm 0.73}$	$\mathbf{95.38 \pm 0.62}$	$\mathbf{94.61 \pm 1.66}$
	{1, 2, 3}	✗	97.69 ± 0.63	95.64 ± 0.36	93.08 ± 1.09	89.23 ± 1.66	82.31 ± 1.26
		✓	$\mathbf{98.46 \pm 0.63}$	$\mathbf{97.18 \pm 1.31}$	$\mathbf{95.64 \pm 0.96}$	$\mathbf{92.56 \pm 2.54}$	$\mathbf{88.46 \pm 2.27}$
Hand-written	{1}	✗	98.42 ± 0.51	98.25 ± 0.35	97.92 ± 0.12	97.92 ± 0.12	97.58 ± 0.12
		✓	$\mathbf{99.50 \pm 0.41}$	$\mathbf{99.50 \pm 0.41}$	$\mathbf{99.50 \pm 0.41}$	$\mathbf{99.50 \pm 0.41}$	$\mathbf{99.50 \pm 0.41}$
	{2}	✗	98.17 ± 1.03	97.75 ± 0.54	97.33 ± 0.47	97.00 ± 0.41	96.92 ± 0.42
		✓	$\mathbf{98.83 \pm 0.24}$	$\mathbf{98.50 \pm 0.41}$	$\mathbf{98.67 \pm 0.47}$	$\mathbf{98.67 \pm 0.47}$	$\mathbf{98.67 \pm 0.47}$
	{1, 2}	✗	97.67 ± 0.47	97.25 ± 0.54	96.58 ± 0.51	95.92 ± 0.59	95.67 ± 0.94
		✓	$\mathbf{99.00 \pm 0.00}$	$\mathbf{98.83 \pm 0.24}$	$\mathbf{98.83 \pm 0.24}$	$\mathbf{98.83 \pm 0.24}$	$\mathbf{98.83 \pm 0.24}$
	{1, 3}	✗	98.08 ± 0.12	97.00 ± 1.22	96.33 ± 1.55	95.33 ± 1.55	95.08 ± 1.59
		✓	$\mathbf{99.50 \pm 0.00}$	$\mathbf{99.17 \pm 0.47}$	$\mathbf{98.00 \pm 0.41}$	$\mathbf{97.67 \pm 0.24}$	$\mathbf{95.78 \pm 1.04}$
	{2, 3}	✗	98.17 ± 0.24	96.83 ± 0.47	95.67 ± 0.85	94.75 ± 0.94	94.17 ± 0.85
		✓	$\mathbf{99.00 \pm 0.00}$	$\mathbf{98.67 \pm 0.47}$	$\mathbf{97.67 \pm 0.62}$	$\mathbf{97.33 \pm 1.03}$	$\mathbf{97.33 \pm 1.03}$
	{1, 2, 3}	✗	96.50 ± 1.08	93.58 ± 1.74	90.67 ± 3.09	88.75 ± 3.54	87.58 ± 3.36
		✓	$\mathbf{98.50 \pm 0.71}$	$\mathbf{97.17 \pm 1.18}$	$\mathbf{95.50 \pm 1.22}$	$\mathbf{93.67 \pm 0.85}$	$\mathbf{92.50 \pm 0.82}$
CUB	{1}	✗	91.11 ± 1.04	86.94 ± 2.83	83.61 ± 3.93	80.83 ± 4.14	79.17 ± 3.79
		✓	$\mathbf{93.33 \pm 1.80}$	$\mathbf{90.83 \pm 2.45}$	$\mathbf{87.50 \pm 3.60}$	$\mathbf{85.56 \pm 4.38}$	$\mathbf{81.11 \pm 4.53}$
	{2}	✗	91.11 ± 0.40	91.95 ± 0.39	91.11 ± 0.40	89.72 ± 0.39	88.61 ± 0.79
		✓	$\mathbf{93.61 \pm 1.04}$	$\mathbf{92.78 \pm 1.04}$	$\mathbf{92.50 \pm 1.80}$	$\mathbf{91.67 \pm 2.96}$	$\mathbf{91.39 \pm 3.22}$
	{1, 2}	✗	92.78 ± 1.97	88.61 ± 1.42	85.83 ± 1.80	79.72 ± 2.83	74.17 ± 4.46
		✓	$\mathbf{94.72 \pm 2.19}$	$\mathbf{92.22 \pm 3.75}$	$\mathbf{90.00 \pm 4.46}$	$\mathbf{86.11 \pm 4.10}$	$\mathbf{79.17 \pm 4.91}$
Animal	{1}	✗	86.61 ± 0.20	85.81 ± 0.36	84.82 ± 1.02	83.77 ± 1.29	82.16 ± 2.32
		✓	$\mathbf{87.20 \pm 0.18}$	$\mathbf{87.01 \pm 0.18}$	$\mathbf{86.60 \pm 0.20}$	$\mathbf{86.03 \pm 0.04}$	$\mathbf{85.42 \pm 0.29}$
	{2}	✗	86.33 ± 0.54	85.62 ± 0.61	84.84 ± 0.95	83.04 ± 1.24	81.34 ± 1.73
		✓	$\mathbf{87.04 \pm 0.08}$	$\mathbf{86.64 \pm 0.26}$	$\mathbf{85.95 \pm 0.42}$	$\mathbf{84.78 \pm 0.17}$	$\mathbf{82.71 \pm 0.24}$
	{1, 2}	✗	86.01 ± 0.17	84.80 ± 0.81	83.17 ± 1.65	80.92 ± 2.77	77.42 ± 4.14
		✓	$\mathbf{87.04 \pm 0.42}$	$\mathbf{86.50 \pm 0.15}$	$\mathbf{85.38 \pm 0.34}$	$\mathbf{83.84 \pm 0.65}$	$\mathbf{81.67 \pm 0.75}$
TUANDROMD	{1}	✗	81.14 ± 0.70	78.21 ± 0.92	75.39 ± 1.09	73.21 ± 1.46	71.71 ± 1.26
		✓	$\mathbf{81.99 \pm 1.99}$	$\mathbf{78.79 \pm 2.42}$	$\mathbf{76.37 \pm 2.57}$	$\mathbf{74.36 \pm 2.63}$	$\mathbf{73.19 \pm 2.60}$
	{2}	✗	84.19 ± 0.82	84.43 ± 0.48	84.46 ± 0.35	84.32 ± 0.45	84.21 ± 0.44
		✓	$\mathbf{84.88 \pm 1.62}$	$\mathbf{84.73 \pm 1.89}$	$\mathbf{84.84 \pm 1.76}$	$\mathbf{84.39 \pm 0.89}$	$\mathbf{84.97 \pm 1.52}$
	{1, 2}	✗	83.56 ± 1.23	80.85 ± 1.30	77.85 ± 1.53	75.90 ± 2.07	74.08 ± 2.22
		✓	$\mathbf{83.99 \pm 1.87}$	$\mathbf{81.48 \pm 2.30}$	$\mathbf{78.50 \pm 2.30}$	$\mathbf{76.73 \pm 2.19}$	$\mathbf{75.23 \pm 2.20}$

C.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 (Tab. 3), we show more other results (Tab. 5 6). It is clear that the models equipped with CML are more robust to noise, especially when the noise is much heavier.

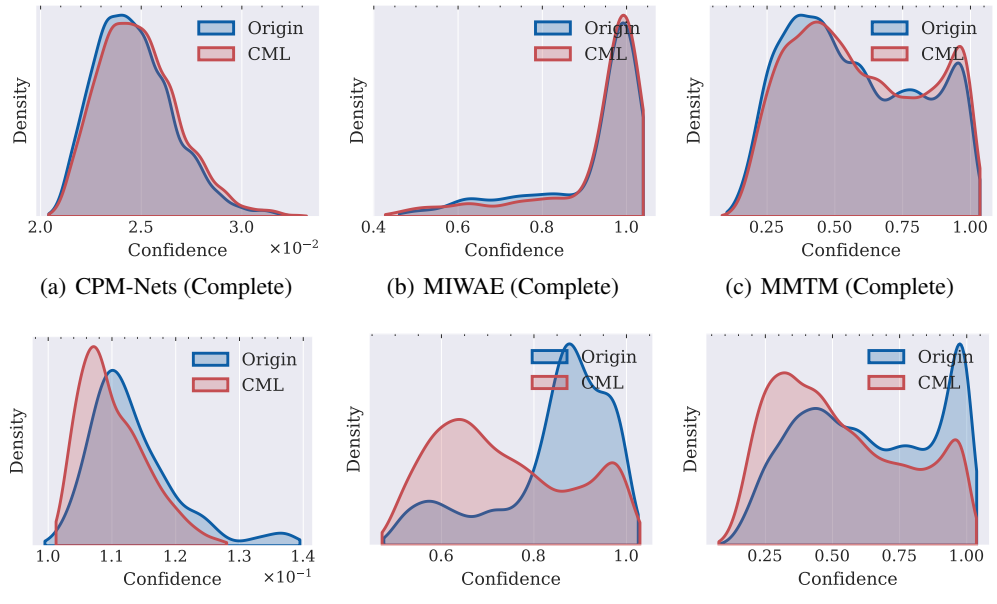
C.4 ADDITIONAL RESULTS FOR ROBUSTNESS ESTIMATION

Table 7: 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

Limited by space, we show the performance of model equipped with CML on YaleB and Handwritten. From Table 7, the classification models equipped with CML consistently outperforms their counterpart validating the rationality of CML principle.

C.5 CONFIDENCE ESTIMATION FOR COMPLETE INPUTS



(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.

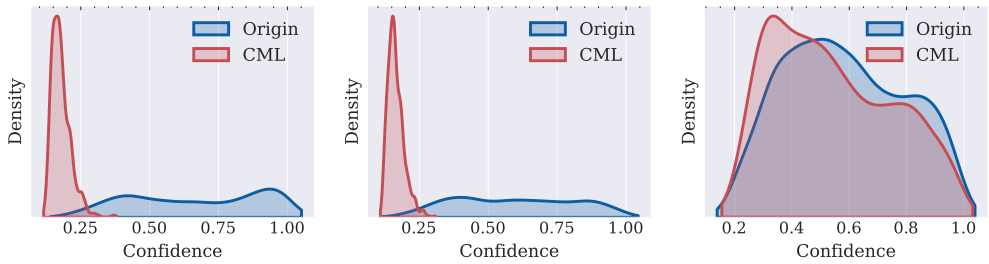
We show the confidence estimation for complete inputs, as shown in Fig. 6, we can find that the confidence estimation of original model and CML model are very similar. To prevent the model from being over-confident when model predicts a wrong prediction, the regularization will not be added when prediction 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 3.2: if we have $\text{VRR}_{CML} < \text{VRR}_{\text{Ori}}$, then we have $\mathbb{E}\left(\text{Conf}_{CML}(\boldsymbol{x}^{(\mathbb{T})})\right) - \mathbb{E}\left(\text{Conf}_{CML}(\boldsymbol{x}^{(\mathbb{S})})\right) \leq \mathbb{E}\left(\text{Conf}_{\text{Ori}}(\boldsymbol{x}^{(\mathbb{T})})\right) - \mathbb{E}\left(\text{Conf}_{\text{Ori}}(\boldsymbol{x}^{(\mathbb{S})})\right)$, then we have:

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

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

C.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.

Forcing the confidence for $\mathbf{x}^{(\mathbb{T})}$ to be smaller than the confidence for $\mathbf{x}^{(\mathbb{S})}$ strictly (Eq. 3) will lead to a very small confidence for $\mathbf{x}^{(\mathbb{T})}$ 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.

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 the imputation-independent model

D.2 CML for the 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.

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 λ of CML, epochs for training the classifier $epoch$
for $e = 1, \ldots, epoch$ **do**
  $\mathbb{S} \leftarrow \mathbb{M}$
  Make the prediction via input $\mathbb{S}$
  $\mathcal{L}^{\text{cl}} \leftarrow \mathcal{L}^{\text{cl}}(\mathbf{x}^{(\mathbb{S})})$
  $\mathcal{L}^{\text{CML}} \leftarrow 0$ **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 $\mathbb{T}$
    $\mathcal{L}^{\text{cl}} \leftarrow \mathcal{L}^{\text{cl}} + \mathcal{L}^{\text{cl}}(\mathbf{x}^{(\mathbb{T})})$
    $\mathcal{L}^{\text{CML}} \leftarrow \mathcal{L}^{\text{CML}} + \max\left(0, \text{Conf}(\mathbf{x}^{(\mathbb{T})}) - \text{Conf}(\mathbf{x}^{(\mathbb{S})}) - \tau\right)$
  end
  $\mathcal{L} = \frac{1}{M}\mathcal{L}^{\text{cl}} + \lambda\mathcal{L}^{\text{CML}}$
  Update the parameters of the classification model with $\mathcal{L}$
end
Return the classifier f_{cl}

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_{re} and classifier f_{cl}, reconstruction loss function $\mathcal{L}^{\text{re}}$, Coefficient λ of CML, epochs for training the reconstruction net $epoch_{\text{re}}$ and classifier $epoch_{\text{cl}}$
for $e_1 = 1, \ldots, epoch_{re}$ **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 $e_2 = 1, \ldots, epoch_{cl}$ **do**
  $\mathbb{S} \leftarrow \mathbb{M}$
  $\mathcal{L}^{\text{CE}} \leftarrow \mathcal{L}^{\text{CE}}(\mathbf{x}^{(\mathbb{S})})$
  $\mathcal{L}^{\text{CML}} \leftarrow 0$
  for $m = M-1, \ldots, 1$ **do**
    Randomly erase a modality of $\mathbb{S}$ and set it as $\mathbb{T}$
    Reconstruct the erased modalities via reconstruction model and add them to $\mathbf{x}^{(\mathbb{T})}$
    Compute the classification loss $\mathcal{L}^{\text{CE}}(\mathbf{x}^{(\mathbb{T})})$ with Cross-Entropy loss function
    $\mathcal{L}^{\text{CE}} \leftarrow \mathcal{L}^{\text{CE}} + \mathcal{L}^{\text{CE}}(\mathbf{x}^{(\mathbb{T})})$
    $\mathcal{L}^{\text{CML}} \leftarrow \mathcal{L}^{\text{CML}} + \max\left(0, \text{Conf}(\mathbf{x}^{(\mathbb{T})}) - \text{Conf}(\mathbf{x}^{(\mathbb{S})}) - \tau\right)$
  end
  $\mathcal{L} = \frac{1}{M}\mathcal{L}^{\text{CE}} + \lambda\mathcal{L}^{\text{CML}}$
  Update the parameters of the classification model with $\mathcal{L}$
end
Return the reconstruction model f_{re} and classifier f_{cl}

E Discussion

E.1 Data-imbalanced

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

CML can improve performance when the data for the training model is data-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 data-imbalanced dataset Animal (data distribution is shown in Fig. 8) validate the effectiveness.

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

Animal is a data-imbalanced real-world dataset, the improvement shows CML can also deal with applications that suffer from data-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 When additional modality is corrupted

◦ **Why the confidence should not decrease even when the additional modality is corrupted?**

Intuitively, the predictive confidence should decrease when some features are corrupted:

$$\text{Conf}(\mathbf{x}) > \text{Conf}(\mathbf{x}_\epsilon), \tag{8}$$

where $\boldsymbol{x}_\epsilon$ indicates the input $\boldsymbol{x}$ is corrupted by the perturbation ϵ. It seems CML is no longer work, and the confidence relationship between the $\mathbf{x}^{(\mathbb{T})}$ and $\mathbf{x}^{(\mathbb{S})}$ should be changed:

$$\text{Conf}(\mathbf{x}^{(\mathbb{T})}) > \text{Conf}(\mathbf{x}^{(\mathbb{S})}). \tag{9}$$

Mathematically, however, Eq. 9 is not the multi-modality representation of Eq. 8, and the Eq. 10 should hold:

$$\text{Conf}(\mathbf{x}^{(\mathbb{S})}) = \text{Conf}(\mathbf{x}^{(\mathbb{T})} \cup \mathbf{x}^{(\complement_{\mathbb{S}}\mathbb{T})}) > \text{Conf}(\mathbf{x}^{(\mathbb{T})} \cup \mathbf{x}_\epsilon^{(\complement_{\mathbb{S}}\mathbb{T})}) \tag{10}$$

Theoretically, the confidence decrease is in terms of $\mathbf{x}^{(\mathbb{S})}$ rather than $\mathbf{x}^{(\mathbb{T})}$ when the additional modalities $\mathbf{x}^{(\complement_{\mathbb{S}}\mathbb{T})}$ are corrupted.

F CML being deployed in advanced multimodal models

MMTM is a SOTA method in multimodal learning 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 multi-modality 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.

Table 8: 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	$\triangle$ 11.10	$\triangle$ 4.53

Table 9: 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	$\triangle$ 0.92	$\triangle$ 1.28	$\triangle$ 0.52	$\triangle$ 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	$\triangle$ 1.29	$\triangledown$ 0.09	$\triangle$ 0.82	$\triangle$ 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	$\triangle$ 1.44	$\triangle$ 1.07	$\triangle$ 0.52	$\triangle$ 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	$\triangle$ 0.78	$\triangle$ 0.97	$\triangle$ 0.57	$\triangle$ 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	$\triangledown$ 0.21	$\triangledown$ 0.13	$\triangledown$ 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	$\triangle$ 0.28	–	$\triangledown$ 0.30	$\triangle$ 0.31

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.

Incomplete multimodal learning Recently, there have been a wide range of research interests in handling missing modalities for multimodal learning, including imputation-independent (Type I) methods (Zhang et al., 2019) and imputation-dependent (Type II) methods (Mattei & Frellsen, 2019; Wu & Goodman, 2018). Imputation-independent methods have no need to reconstruct the miss-

Table 10: 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

ing 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 (i.e., Type I) 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 (i.e., Type II) in multimodal learning, 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.