Predictive Dynamic Fusion

Bing Cao [1][2] **Yinan Xia** [1] **Yi Ding** [1] **Changqing Zhang** [1][2] **Qinghua Hu** [1][2]

Abstract

Multimodal fusion is crucial in joint decision-making systems for rendering holistic judgments. Since multimodal data changes in open environments, dynamic fusion has emerged and achieved remarkable progress in numerous applications. However, most existing dynamic multimodal fusion methods lack theoretical guarantees and easily fall into suboptimal problems, yielding unreliability and instability. To address this issue, we propose a Predictive Dynamic Fusion (PDF) framework for multimodal learning. We proceed to reveal the multimodal fusion from a generalization perspective and theoretically derive the predictable Collaborative Belief (Co-Belief) with Mono- and Holo-Confidence, which provably reduces the upper bound of generalization error. Accordingly, we further propose a relative calibration strategy to calibrate the predicted Co-Belief for potential uncertainty. Extensive experiments on multiple benchmarks confirm our superiority. Our code is available at `https://github.com/Yinan-Xia/PDF`.

1. Introduction

Many decision-making challenges in real-world applications, such as autonomous driving (Cui et al., 2019; Feng et al., 2020), clinical diagnosis (Perrin et al., 2009; Tempany et al., 2015), and sentiment analysis (Soleymani et al., 2017; Zadeh et al., 2017), are fundamentally based on multimodal data (Kiela et al., 2019). To fully capture complementary perceptions, multimodal fusion emerges as a promising learning paradigm that presents an opportunity to integrate all available modalities and achieve enhanced performance. Despite these advances, experiments have shown that tradi-

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China [2]Tianjin Key Lab of Machine Learning, Tianjin, China. Correspondence to: Changqing Zhang <zhangchangqing@tju.edu.cn>, Qinghua Hu <huqinghua@tju.edu.cn>.
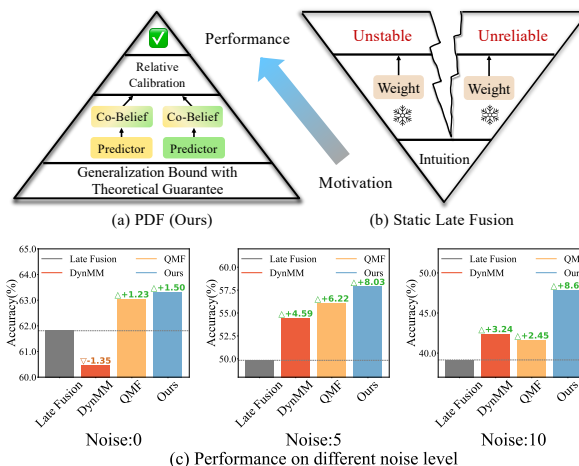
Figure 1. Our PDF *v.s.* other fusion methods. We derive from the upper bound of generalization error and predict the Co-Belief for each modality with a theoretical guarantee. The relative calibration calibrates potential uncertainty for more reliable learning. Experiments on different noise levels validate our superiority.

tional fusion techniques have largely overlooked the dynamically changing quality of multimodal data (Natarajan et al., 2012; Pérez-Rúa et al., 2019; Yan et al., 2004). In reality, the data quality of different modalities and their inherent relationships often vary with the open environment. Numerous studies (Xue & Marculescu, 2023) empirically recognized that multimodal learning sometimes falls to depending on partial modalities, even a single modality, instead of multimodal data, especially with modality imbalance (Wang et al., 2020; Peng et al., 2022) or high noise (Huang et al., 2021c; Scheunders & De Backer, 2007). Therefore, dynamic multimodal learning becomes a key cue for robust fusion. Some recent works theoretically proved multimodal learning models do not always outperform their unimodal counterparts, encountering limited data volumes (Huang et al., 2021b). This indicates that the dynamic relationship between multimodal data is not a free lunch.

Intuitively, fusing information from multimodal data by using the overall quality estimation of each modality is reasonable. However, the data quality estimation is not always reliable due to the unimodal uncertainty and the changing relative reliability of multimodal systems (Ma et al., 2023). We empirically identify that the dominance of each modality

is dynamically changing in open environments. On this basis, one fundamental challenge to reliable multimodal learning is how to precisely estimate the contribution of each modality to the multimodal systems (Zhang et al., 2023). However, existing multimodal dynamic fusion techniques mainly focus on addressing this problem by exploring dynamic network architecture (Xue & Marculescu, 2023) or estimating the modality's quality by uncertainty (Han et al., 2022b), which commonly lack theoretical guarantees, resulting in unsatisfactory fusion performance.

To solve this problem, we revisited the relationship between modality fusion weights and losses. Deriving from the upper bound of generalization error (Theorem 3.5 in Mohri et al. 2018), we reveal that the key to reducing generalization error bound lies in the negative covariance between fusion weight and current modality loss as well as the positive covariance between fusion weight and other's modality loss, which implies that fusion weights in the multimodal system should not solely consider the unimodal but is also enforced to integrate other modalities' statuses. With this finding, a natural idea is to employ the loss value of each modality to perform multimodal fusion. However, directly predicting the loss value is unstable as the loss is minimized when converged (see Section 5.3.3). In the setting of multimodal classification with cross-entropy loss, we transform the prediction of loss value into the confidence of true class label, while satisfying the correlation derived from the generalization error. The motivation is based on a natural intuition, i.e., the probability of true class and loss is negative.

To this end, we offer a new perspective on the theoretical foundation of multimodal fusion and propose a Predictive Dynamic Fusion (PDF) framework, which is effective in reducing the upper bound of generalization error and significantly improving multimodal reliability and stability. As shown in Figure 1 (a), to be specific, PDF predicts the Collaborative Belief (Co-Belief) with Mono-Confidence and Holo-Confidence for each modality. The Mono- and Holo-Confidences derive from the intra-modal negative and inter-modal positive covariance between fusion weight and loss function, respectively. In addition, we empirically identify the changing data quality in open environments, which leads to inevitable prediction uncertainty. To handle this issue, we further propose a relative calibration to calibrate the predicted Co-Belief from the perspective of a multimodal system, which implies that the relative dominance of each modality should change dynamically as the quality of other modalities changes, rather than being static. Experiments demonstrated that our method has strong generalization capabilities, achieving superior results on multiple datasets. Overall, our contributions can be summarized as follows:

- This paper provides an intuitive and rigorous multimodal fusion paradigm from the perspective of generalization error. Under theoretical analysis, we derive

a new Predictive Dynamic Fusion (PDF) framework based on the covariance of the fusion weight and loss function. This offers theoretical guarantees to reduce the upper bound of generalization error in decision-level multimodal fusion.

- We propose to transform the loss prediction to a more robust Collaborative Belief (Co-Belief) prediction, which naturally satisfies the covariance relationship to reduce the upper bound of generalization error without additional computational cost, and significantly enhance the prediction stability.

- We develop a relative calibration strategy to calibrate the potential prediction uncertainty and reveal the relative dominance in dynamic multimodal systems. Extensive experiments validate our theoretical analysis and superior performance.

2. Related Works

Multimodal fusion is a fundamental problem in multimodal learning (Atrey et al., 2010; Cao et al., 2023; Zhu et al., 2024). Existing methods can be mainly categorized into early fusion (Nefian et al., 2002), middle fusion (Natarajan et al., 2012), and late fusion (Snoek et al., 2005; Wang et al., 2019b). Early fusion (Ayache et al., 2007) directly combines various modalities at the data level, often merging multimodal data through concatenation. Middle fusion (Han et al., 2022a; Wang et al., 2019a) is widely used in multimodal learning, which mainly fuses multimodal data at the feature level. Late fusion (Zhang et al., 2023) usually integrates multimodal data in the semantic space, which can be further grouped into naive fusion (Liu et al., 2018), learnable classifier fusion (Xue & Marculescu, 2023), and confidence-based fusion (Han et al., 2022b).

Uncertainty estimation is crucial for improving the model's interpretability, accuracy, and robustness, especially for multimodal systems. Many efforts (Neal, 2012; Gal & Ghahramani, 2016) have been made on this issue. Bayesian Neural Networks (BNNs) (Denker & LeCun, 1990; Mackay, 1992) use probability distributions, rather than single values, to represent the weights in neural networks. Deep ensemble methods (Lakshminarayanan et al., 2017; Amini et al., 2020) typically train multiple models and aggregate their predictions, then estimate the uncertainty through prediction variances. Dempster-Shafer's theory extends Bayesian to subjective probabilities, offering a robust model for handling epistemic uncertainty (Dempster, 1968). Energy Score (Liu et al., 2020) is promising in estimating uncertainty. The pioneering work, QMF (Zhang et al., 2023) explores generalization error and uncertainty-aware weighting to perform robust fusion. Gradient-based uncertainty (Lee & AlRegib, 2020) uses backward propagation gradients to quantify uncertainty. Essentially, it evaluates the output uniformity.
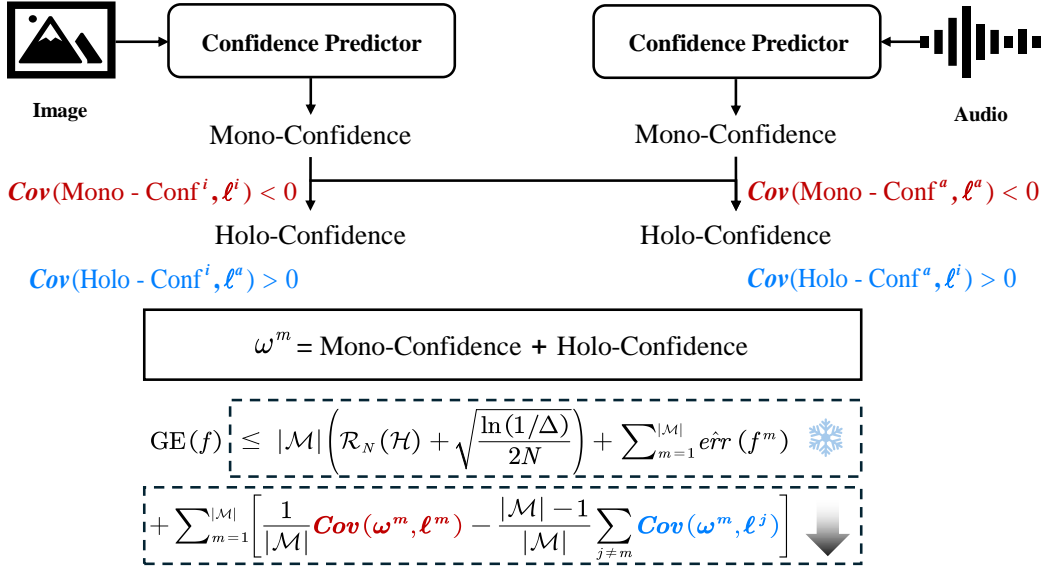
$$\text{GE}(f) \leq |\mathcal{M}| \left(\mathcal{R}_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\Delta)}{2N}} \right) + \sum_{m=1}^{|\mathcal{M}|} e\hat{r}r(f^m)$$

$$+ \sum_{m=1}^{|\mathcal{M}|} \left[\frac{1}{|\mathcal{M}|} \boldsymbol{Cov}(\boldsymbol{\omega}^m, \boldsymbol{\ell}^m) - \frac{|\mathcal{M}|-1}{|\mathcal{M}|} \sum_{j \neq m} \boldsymbol{Cov}(\boldsymbol{\omega}^m, \boldsymbol{\ell}^j) \right]$$

Figure 2. We use confidence predictors to predict the Mono-Confidence of each modality, where the confidence is negatively correlated with the loss of the corresponding modality theoretically. Taking into account the Mono-Confidence of other modalities, we further obtain the Holo-Confidence, where the confidence is positively correlated with the loss of other modalities. By combining Mono-Confidence and Holo-Confidence, we obtain the Co-Belief, which is calibrated as fusion weight to achieve a reduction in the generalization error bounds.

3. Theory

In this section, we first clarify the basic settings and formulas in multimodal fusion. Next, we revisit the formula for generalization error bounds and establish its connection with fusion weights, revealing the theoretical guarantee for reducing the upper bound of generalization error. Finally, we propose a predictable dynamic fusion framework that satisfies the above theoretical analysis.

3.1. Basic Setting

Given multimodal tasks, we define $\mathcal{M}$ as the set of modalities, thus $|\mathcal{M}|$ is the cardinality of $\mathcal{M}$. We denote our training dataset as $\mathcal{D}_{train} = \{x_i, y_i\}_{i=1}^N \in \mathcal{X} \times \mathcal{Y}$, where N is the sample size of $\mathcal{D}_{train}$, $x_i = \{x_i^m\}_{m=1}^{|\mathcal{M}|}$ has $|\mathcal{M}|$ modalities, and $y_i \in \mathcal{Y}$ is the corresponding label. We aim to design a predictable fusion weight ω for each modality and achieve a robust multimodal fusion. The uni-modal projection function $f^m : \mathcal{X} \rightarrow \mathcal{Y}$ is trained as the fusion weight ω^m dynamically adjusting during training, where $m \in \mathcal{M}$. The decision-level multimodal fusion is as:

$$f(x) = \sum_{m=1}^{|\mathcal{M}|} \omega^m \cdot f^m(x^m). \quad (1)$$

3.2. Generalization Error Upper Bound

Generalization Error Upper Bound (GEB) is an important concept in machine learning, referring to an upper bound

on the performance of a model on unknown data (Zhang et al., 2023). Typically, the smaller the upper bound of the generalization error, the better the model's generalization ability, i.e., the better it performs on unknown joint distribution. For binary classification, the Generalization Error (GE) of a model f can be defined as:

$$\text{GE}(f) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(f(x), y)], \quad (2)$$

where ℓ is a convex logistic loss function, $\mathcal{D}$ is an unknown dataset. By Rademacher complexity theory (Theorem 3.5 in Mohri et al. 2018), we delve into the essence of GEB in multimodal fusion and obtain the following theorem. The full proof of Theorem 3.1 is given in Appendix A.1.

Theorem 3.1. *(Generalization Error Upper Bound in Multimodal System). Let $e\hat{r}r(f^m)$ denotes the empirical errors of the m-th modality on $\mathcal{D}_{train} = \{x_i, y_i\}_{i=1}^N$, and $\mathcal{H}$ is hypothesis set i.e., $\mathcal{H} : \mathcal{X} \rightarrow \{-1, +1\}$, where $f \in \mathcal{H}$. $\mathcal{R}_N(\mathcal{H})$ is the Rademacher complexities (Theorem 3.5 in Mohri et al. 2018). We holds with a confidence level of $1 - \Delta$ $(0 < \Delta < 1)$:*

$$\text{GE}(f) \leq |\mathcal{M}| \left(\mathcal{R}_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\Delta)}{2N}} \right) + \sum_{m=1}^{|\mathcal{M}|} e\hat{r}r(f^m)$$

$$+ \sum_{m=1}^{|\mathcal{M}|} \left[\frac{1}{|\mathcal{M}|} \underbrace{Cov(\omega^m, \ell^m)}_{Mono\text{-}Covariance} - \frac{|\mathcal{M}|-1}{|\mathcal{M}|} \sum_{j \neq m} \underbrace{Cov(\omega^m, \ell^j)}_{Holo\text{-}Covariance} \right],$$

$$(3)$$

where $Cov(\omega^m, \ell^m)$ is the covariance between fusion

weight and loss of m-th modality, and $Cov(\omega^m, \ell^j)$ is the cross-modal covariance. Note that the empirical errors $e\hat{r}r(f^m)$ and the Rademacher complexities $\mathcal{R}_N(\mathcal{H})$ remain constant when optimizing over the same model. Therefore, the key to achieving a lower GEB lies in ensuring that $Cov(\omega^m, \ell^m) < 0$ and $Cov(\omega^m, \ell^j) > 0, \forall j \neq m$. Thus, to reduce the multimodal fusion model's GEB, we can draw the following corollaries:

Corollary 3.2. *A negative correlation should exist between a modality's weight and its loss.*

Corollary 3.3. *A positive correlation is desirable between a modality's weight and the losses of the other modalities.*

3.3. Collaborative Belief

3.3.1. MONO-CONFIDENCE

To fulfill Corollary 3.2, an intuitive strategy is to predict the loss of each modality and utilize the predicted loss to formulate the weight, thereby establishing a negative correlation explicitly and directly. Nevertheless, employing loss as a fusion weight for modalities presents significant challenges. Notably, as the loss minimizes during the training process, even marginal biases can induce substantial perturbations. This sensitivity to small errors in loss estimation may compromise the stability and effectiveness of the weight. Also, the loss value may range from zero to positive infinity, making its precise prediction quite challenging. To mitigate these challenges, we propose substituting loss with the probability ($p_{true} \in [0,1]$) of the true class label, which is inversely related to the loss as denoted by $\ell = -\log p_{true}$ (Full derivation of the relationship between loss and p_{true} is provided in the Appendix A.2), provides a more stable and interpretable basis for weight computation.

Owing to the negative correlation between loss and p_{true}, we consider using the p_{true} as the weight for multimodal fusion to fulfill Corollary 3.2. By analyzing the properties of p_{true}, we identify that it reflects the confidence of modality, some works (Corbière et al., 2019) have articulated this. Using p_{true} as a fusion weight not only helps in lowering the upper bound of generalization error but also provides a theoretical guarantee for dynamic multimodal fusion. Since the predictable p_{true} solely considers the current modality's confidence, we define it as Mono-Confidence:

$$\text{Mono-Conf}^m = \hat{p}_{true}^m, \tag{4}$$

where $\hat{p}_{true}^m$ is the prediction of p_{true}^m as there is no ground-true label in the test phase. The detailed implementation of the prediction is given in Appendix C.6

3.3.2. HOLO-CONFIDENCE

Recalling Corollary 3.3, an instinctive approach to naturally achieve $Cov(\omega^m, \ell^j) > 0, \forall j \neq m$ is using the losses of other modalities as the weight. Thus, we consider constructing the weight by using the sum of the losses from other modalities as the weight for the given modality. Basing the property of $\ell = -\log p_{true}$, we replaced ℓ with p_{true}. Thus, we define this term as Holo-Confidence because of the cross-modal interaction of p_{true}:

$$
\begin{aligned}
\text{Holo-Conf}^m &= \frac{\sum_{j \neq m}^{|\mathcal{M}|} \hat{\ell}^j}{\sum_{i=1}^{|\mathcal{M}|} \hat{\ell}^i} \\
&= \frac{\log \prod_{j \neq m} \hat{p}_{ture}^j}{\log \prod_{i=1}^{|\mathcal{M}|} \hat{p}_{true}^i}, \tag{5}
\end{aligned}
$$

where $\hat{\ell}^i$ and $\hat{\ell}^j$ is the prediction of ℓ^i and ℓ^j. Our proposed Holo-Confidence also fulfills Corollary 3.3, the full proof is given in the Appendix A.3.

3.3.3. CO-BELIEF

Since Mono-Confidence and Holo-Confidence facilitate the collaborative interaction among modalities, to fulfill Corollary 3.2 and 3.3 simultaneously, we define a collaborative belief (Co-Belief) as a linear combination of the predictable Mono-Confidence and Holo-Confidence, which can be taken as the final fusion weight.

$$\text{Co-Belief}^m = \text{Mono-Conf}^m + \text{Holo-Conf}^m. \tag{6}$$

Noting that the Co-Belief meets Corollary 3.2 and 3.3 simultaneously, and is better than the individual term to represent the weight. The proof is shown in Appendix A.4.

4. Method

To achieve a reliable prediction, we propose a relative calibration strategy to calibrate the predicted Co-Belief to handle inevitable uncertainty. With this reliable prediction, we coined our multimodal fusion framework as **P**redictive **D**ynamic **F**usion (PDF).

Note that the data quality usually dynamically changes in open environments, leading to inevitable uncertainty for the prediction. To decrease the potential uncertainty of Co-Belief in complex scenarios, we further propose a Relative Calibration (RC) to calibrate the predicted Co-Belief from the perspective of a multimodal system. This implies that the relative dominance of each modality should change dynamically as the quality of other modalities changes, rather than being static.

Firstly, we define the **D**istribution **U**niformity DU^m of m-th modality in multimodal system as,

$$\text{DU}^m = \frac{1}{C} \sum_{i=1}^{C} |Softmax(f^m(x^m))_i - \mu|, \tag{7}$$

where C is the class number, μ is the mean of probability, and it holds $\mu = \frac{1}{C}$. The distribution of probabilities after softmax offers critical insights into a model's uncertainty: A uniform distribution typically suggests high uncertainty, whereas a peaked distribution implies low uncertainty in predictions (Huang et al., 2021a). We compared it with other uncertainty estimation methods in Appendix B.2.

Considering the changing environment, the uncertainty of different modalities in a multimodal system should be relative, i.e., the uncertainty of each modality should change dynamically as the uncertainty of other modalities changes. One modality should dynamically perceive the changes of other modalities and modify its relative contribution to the multimodal system. Thus, we introduce a **R**elative **C**alibration (RC) to calibrate the relative uncertainty of each modality. The relative calibration for m-th modality can be formulated as follows (in a scenario with two modalities, denoted as $m, n \in \mathcal{M}$):

$$\mathrm{RC}^m = \frac{\mathrm{DU}^m}{\mathrm{DU}^n} = \frac{\sum\limits_{i=1}^{C} |Softmax(f^m(x^m))_i - \mu|}{\sum\limits_{i=1}^{C} |Softmax(f^n(x^n))_i - \mu|}. \quad (8)$$

The definition of RC^m when $|\mathcal{M}| > 2$ is given in the Appendix A.5.

Considering real-world factors, RC^m works with an asymmetric form to further calibrate the Co-Belief. Specifically, we postulate that the modality with a $\mathrm{RC}^m < 1$ possess greater uncertainty, and it tends to produce relatively unreliable predictions for $\hat{p}_{true}^m$ (Gawlikowski et al., 2023), thus the corresponding Co-Belief has potential risks in accuracy. Hence, we reduce the contribution of such a modality by multiplying its predicted Co-Belief with RC^m ($\mathrm{RC}^m < 1$). Conversely, modalities with a $\mathrm{RC}^m > 1$ are deemed to have less uncertainty and accurate Co-Belief, thus the contribution of these modalities can be maintained to reduce optimization difficulty. Based on this, the asymmetric calibration term is defined as:

$$k^m = \begin{cases} \mathrm{RC}^m = \frac{\mathrm{DU}^m}{\mathrm{DU}^n} & \text{if } \mathrm{DU}^m < \mathrm{DU}^n, \\ 1 & \text{otherwise.} \end{cases} \quad (9)$$

We calibrate the Co-Belief of the m-th modality with our asymmetric calibration strategy and obtain the Calibrated Co-Belief (CCB) as:

$$\mathrm{CCB}^m = (\text{Co-Belief}^m) \cdot k^m. \quad (10)$$

We use the each modality's CCB as its fusion weight in a multimodal system,

$$f(x) = \sum_{m=1}^{|\mathcal{M}|} \omega^m \cdot f^m(x^m)$$
$$= \sum_{m=1}^{|\mathcal{M}|} Softmax(\mathrm{CCB}^m) \cdot f^m(x^m). \quad (11)$$

The loss functions of our PDF framework are given in Appendix C.6.

5. Experiments

5.1. Setup

Datasets. We evaluate the proposed method across various multimodal classification tasks, including ⋄ *Image-text classification*: The UPMC FOOD101 dataset (Wang et al., 2015) contains noisy images and texts obtained in uncontrolled environments containing about 100,000 recipes for a total of 101 food categories. MVSA (Niu et al., 2016) is a sentiment analysis dataset that collects sentiment data for matched pairs of users' texts and images; ⋄ *Scenes recognition*: NYU Depth V2 (Silberman et al., 2012) is an indoor scenes dataset, both the RGB and Depth Cameras recorded the image-pairs; ⋄ *Emotion recognition*: CREMA-D (Cao et al., 2014) is an audio-visual dataset designed for recognizing multi-modal emotion, demonstrating various basic emotional states (happy, sad, anger, fear, disgust, and neutral) through spoken sentences. ⋄ *Face recognition*: PIE (Sim et al., 2003) is a pose, illumination, and expression database of over 40,000 facial images of 68 people.

Evaluation metrics. We report the average and worst accuracies in the presence of Gaussian noise (for image and audio modalities) and blank noise (for text modality), in accordance with prior studies (Zhang et al., 2023; Han et al., 2022b; Xie et al., 2017; Ma et al., 2021). To mitigate the impact of randomness, we replicate to evaluate our model using five distinct seeds. We also defined two new metrics, **A**ggregate **C**ovariance (AC) and **GEB D**ecreasing **P**roportion (GDP), to quantify the capability of fusion strategy to reduce generalization error upper bound, i.e. the generalization ability of the model with certain fusion strategy. The specific definitions are given in Appendix C.3.

Competing methods. In our experiments, we compare our method with established fusion techniques, including late fusion and concatenation-based fusion, which are static, as well as with dynamic fusion approaches, including TMC (Han et al., 2022b), QMF (Zhang et al., 2023) and DynMM (Xue & Marculescu, 2023). We also establish unimodal baselines for comparison: RGB and depth for scene recognition; text and image for image-text classification; visual and audio for emotion recognition.

Table 1. We add Gaussian noise on 50% modalities and ϵ presents the noise degree. This table reported the average and worst classification accuracies of our method and the competing methods on MVSA, FOOD101, NYU Depth V2, and CREMA-D datasets. The method marked with * was replicated by ourselves, while the rest of the results are sourced from (Zhang et al., 2023). Full results with standard deviation are reported in Table 9.

DATASET	METHOD	$\epsilon = 0.0$		$\epsilon = 5.0$		$\epsilon = 10.0$	
		AVG.	WORST.	AVG.	WORST.	AVG.	WORST.
MVSA	IMG	64.12	62.04	49.36	45.67	45.00	39.31
	TEXT	75.61	74.76	69.50	65.70	47.41	45.86
	CONCAT	65.59	64.74	50.70	44.70	46.12	41.81
	LATE FUSION	76.88	74.76	63.46	58.57	55.16	47.78
	TMC	74.88	71.10	66.72	60.12	60.36	53.37
	QMF	78.07	76.30	73.85	71.10	61.28	57.61
	DYNMM*	79.07	78.23	67.96	65.51	59.21	56.65
	OURS	**79.94**	**78.42**	**74.40**	**72.64**	**63.09**	**60.31**
UMPC FOOD 101	IMG	64.62	64.22	34.72	34.19	33.03	32.67
	TEXT	86.46	86.42	67.38	67.19	43.88	43.56
	CONCAT	88.20	87.81	61.10	59.25	49.86	47.79
	LATE FUSION	90.69	90.58	68.49	65.05	58.00	55.77
	TMC	89.86	89.80	73.92	73.64	61.37	61.10
	QMF	92.92	92.72	76.03	74.68	62.21	61.76
	DYNMM*	92.59	92.50	74.74	74.35	59.68	59.22
	OURS	**93.32**	**92.84**	**76.47**	**76.09**	**62.83**	**62.03**
NYU DEPTH V2	RGB	63.30	62.54	53.12	50.31	45.46	42.20
	DEPTH	62.65	61.01	50.95	42.81	44.13	35.93
	CONCAT*	69.88	69.11	63.82	61.47	60.03	55.66
	LATE FUSION*	70.03	68.65	64.37	63.30	60.55	57.95
	TMC*	70.40	70.03	59.33	55.51	50.61	45.41
	QMF*	69.54	68.65	64.10	62.54	60.18	58.41
	DYNMM*	65.50	64.99	54.31	52.14	46.79	45.26
	OURS	**71.37**	**70.18**	**65.72**	**63.91**	**62.56**	**60.25**
CREMA-D	VISUAL*	43.60	40.05	32.52	28.49	30.17	28.09
	AUDIO*	58.67	57.39	54.66	50.67	43.01	35.35
	CONCAT*	61.56	59.95	52.33	45.16	41.01	31.59
	LATE FUSION*	61.81	57.39	49.84	39.92	39.15	29.90
	TMC*	59.15	56.18	54.42	45.16	46.79	37.63
	QMF*	63.04	60.75	56.06	51.75	41.60	35.89
	DYNMM*	60.46	59.81	54.43	52.82	42.39	41.26
	OURS	**63.31**	**61.69**	**57.85**	**54.17**	**47.84**	**44.62**

Implementation details. The network was trained for 100 epochs utilizing the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 0.01, dropout rate of 0.1, and a batch size of 16. All the experiments were conducted on an NVIDIA A6000 GPU, using PyTorch with default parameters for all methods. More details are provided in Appendix C.4.

5.2. Questions to be Verified

We conducted a series of experiments to investigate our effectiveness and rationale thoroughly. The experiments were designed to address four primary questions:

- *Does our proposed method have better generalization ability than its counterparts?*

In Section 3.2, we conducted a theoretical analysis to demonstrate that our method can effectively lower the upper bound of generalization error, which is evidenced in its performance and robustness to noisy data. To empirically substantiate that our method reduces the generalization error upper bound, we carried out comparative experiments across five datasets under diverse noise conditions.

- *Does our PDF framework really work?*

We performed an ablation study to verify the effectiveness of each component of our framework. Additionally, we visualized the capability of Mono-Confidence, Co-Belief, and Calibrated Co-Belief in reducing the generalization error upper bound.
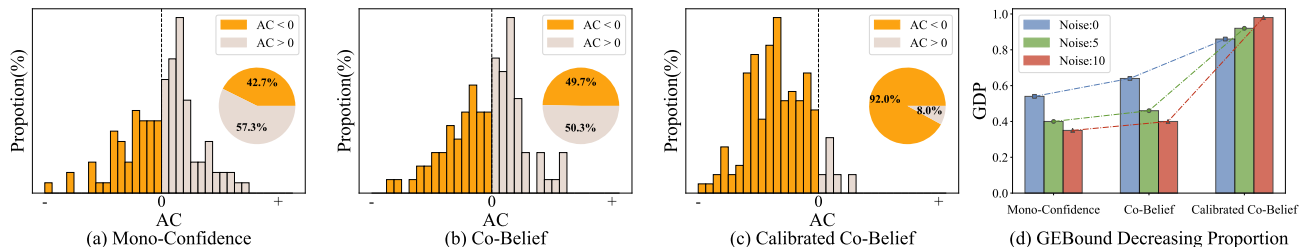
Figure 3. We evaluated the effectiveness of Mono-Confidence, Co-Belief, and Calibrated Co-Belief as fusion weights on the NYU Depth V2 dataset to determine their effectiveness in minimizing the generalization error upper bound. The yellow part of the pie chart in Figure 3 (a), (b), or (c) illustrates the Generalization Error Bound Decreasing Proportion (GDP) for each weight form under varying noises (0, 5, and 10). As depicted in Figure 3 (c), the Calibrated Co-Belief attains the highest GDP, leading to the best generalization. Figure 3 (d) presents the GDP across diverse fusion strategies and noise intensities. More details are given in Appendix D.1.

- *Why do we predict p_{true} instead of loss?*

  In Section 3.3, we choose to establish the correlation between weights and loss by predicting p_{true}. We analyzed the distribution of p_{true}, and by examining the relationship between p_{true} and loss, we identified the challenge in predicting loss. Eventually, we compared the performance of the two prediction methods, validating the advantages of p_{true}.

- *Why relative calibration is effective and reliable?*

  To further investigate the effect of relative calibration, we conducted experiments to explore the relative uncertainty when data quality changes by adding noise. We also compared the efficacy of DU with that of other uncertainty estimation methods.

5.3. Results

5.3.1. GENERALIZATION ABILITY

Our method improves the model's generalization compared to the competing approaches, as shown in Table 1. Notably, as the noise intensity increases, the advantages of our method become increasingly highlighted, reinforcing its superior generalization potential. It is especially commendable that our approach consistently realizes state-of-the-art performance on all the datasets against the competing methods. Full results under different noises (Gaussian noise and Salt-Pepper noise) with standard deviation are shown in Appendix D.5. Additionally, we also conducted experiments on PIE dataset with 3 modalities in Appendix D.4.

To further validate the efficacy of our approach, we incorporated time-varying noise into the CREMA-D dataset to emulate real-world scenarios. Specifically, we introduced noise with varying frequencies to synthesize noisy speech data, and the intensity of noise added to image frames also varied over time. As depicted in Table 3, our PDF demonstrated exceptional generalization prowess, even amidst the influence of time-varying noise interference.

Table 2. Ablation study on MVSA to verify the effectiveness of Mono-Confidence (MC), Holo-Confidence (HC), and relative calibration (RC) as well as the complete model.

MC	HC	RC	$\epsilon = 0$		$\epsilon = 5$		$\epsilon = 10$	
			AVG.	WORST.	AVG.	WORST.	AVG.	WORST.
$\checkmark$			79.43	78.23	72.57	69.56	60.84	55.11
	$\checkmark$		79.28	78.23	72.57	69.94	61.22	55.68
		$\checkmark$	79.11	77.84	71.30	63.97	60.11	50.29
$\checkmark$	$\checkmark$		79.92	**79.00**	72.83	70.13	60.97	55.49
	$\checkmark$	$\checkmark$	79.06	78.23	73.09	71.10	62.11	58.57
$\checkmark$		$\checkmark$	79.62	78.23	73.12	70.91	62.13	57.61
$\checkmark$	$\checkmark$	$\checkmark$	**79.94**	78.42	**74.40**	**72.64**	**63.09**	**60.31**

Table 3. Comparison between our PDF and other competing methods on CREMA-D dataset with added time-varying noise.

METHOD	AVG.	WORST.
CONCAT	60.20	58.60
LATE FUSION	60.80	57.39
TMC	57.17	54.03
QMF	61.20	59.40
DYNMM	57.90	57.53
OURS	**61.40**	**60.34**

5.3.2. ABLATION STUDY

We conducted ablated experiments on Mono-Confidence, Holo-Confidence, and Relative Calibration. The mean and worst accuracy on the MVSA dataset across different noise intensities are reported. The results, presented in Table 2, reveal that the model with Calibrated Co-Belief attains the best robustness and generalization.

Lower generalization error upper bound. We applied the proposed Generalization Error Bound Decreasing Proportion (GDP) to measure the capacity of a fusion strategy
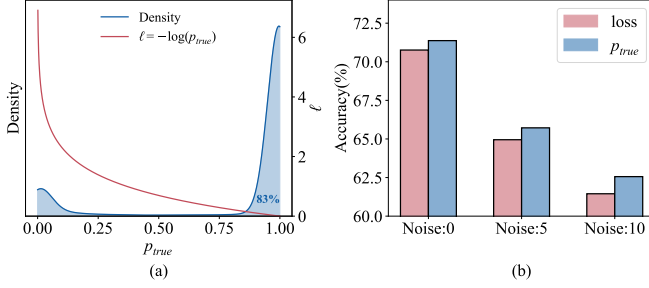
Figure 4. We present the true distribution of p_{true} for the samples in UPMC Food 101, according to the blue area in (a), while the red line in (a) is the function curve of loss corresponding to p_{true}. In (b), we reported the performance of two prediction methods in various noise conditions. It's obvious that predicting p_{true} yields better performance.

Table 4. Comparison with traditional uncertainty on MVSA.

UNCERTAINTY	$\epsilon = 0$	$\epsilon = 5$	$\epsilon = 10$
ENERGY	79.14	73.22	61.72
ENTROPY	78.88	72.07	60.94
EVIDENCE	79.17	73.28	61.87
MCP	79.48	73.22	61.93
OURS (DU)	**79.94**	**74.40**	**63.09**

for reducing the Generalization Error Upper Bound (GEB). Separately, we compare the GDP metrics of models that utilize Mono-Conf (Equation (4)), Co-Belief (Equation (5)), and CCB (Equation (10)) as fusion weights. Specifically, for each fusion strategy, we trained 50 models with distinct random seeds and reported the GDP of these models under varying noises (0, 5, and 10) during testing. We took Mono-Confidence, Co-Belief, and CCB as distinct fusion weights for the model and depicted the AC (defined in Equation (23)) distribution in Figure 3 (a), (b), and (c). The proportion of AC < 0, i.e. GDP, signifies the fusion strategy's potential to reduce the model's GEB. Figure 3 (d) displays the GDP of different fusion strategies under diverse noisy conditions. The proportion of Mono-Confidence that reduces GEB without noise is greater than 50%, and Holo-Confidence further increases the model's GDP, CCB endows the model with the best generalization capabilities in dynamic environments. These experiments validate our effectiveness in lowering the generalization error upper bound.

5.3.3. PREDICTING p_{true} IS MORE FEASIBLE

Although the most direct method to establish a relationship between weight and loss is to use the predicted loss as weight, it encountered difficulties in practice. As shown in Figure 4 (a), we display the sample distribution of p_{true} and the corresponding loss function curve. We observed that ap-
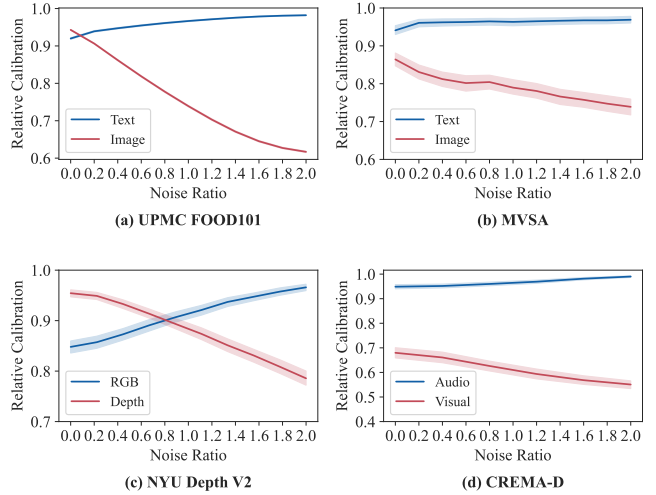

Figure 5. Relative Calibration (RC) can detect noise variations within the current modality as well as in other modalities. The noise ratio denotes the ratio of the noises added to the two modalities, representing the relative exposure of the two modalities to noise. We maintained a fixed noise level for the modality denoted by the blue line in the figure.

proximately 83% of p_{true} values fall into the range of 0.8 to 1, while the corresponding loss value ranges between 0 and 0.097, making it challenging to predict loss accurately. Employing the same methodology, we conducted comparative experiments on predicting the loss and the p_{true}, Figure 4 (b) shows that predicting p_{true} consistently outperforms predicting loss under various noise conditions. This experiment validates the effectiveness of our prediction strategy.

5.3.4. EFFECTIVENESS OF RELATIVE CALIBRATION

Relative calibration reflects the quality of modality. We conducted experiments across four datasets with two modalities to explore the responsiveness of Relative Calibration (RC) to modality quality. We changed the data quality by varying noise levels. Specifically, we alter the degree of noise added to it and fix another modality's noise level. More details of adding noise are given in Appendix C.5.

The noise ratio denotes the ratio of the noises added to the two modalities, representing the relative exposure of the two modalities to noise. As illustrated in Figure 5, the RC value of the noisy modality declines with increased noise, while the RC of the modality with fixed noise enhances or maintains as the noise level of the other modality escalates. This indicates that RC is adept at discerning the quality of both its modality and the others. These findings corroborate the conceptual expectation of multimodal data quality relativity and emphasize the dynamism and interpretability of RC.

Comparison with traditional uncertainty. We compared the proposed DU with conventional uncertainty estimation

methods. For a fair comparison, we substituted our proposed DU in RC with alternative uncertainties, namely energy-based uncertainty (Liu et al., 2020), entropy (Shannon, 1948), uncertainty in Dempster–Shafer Theory (DST) (Sensoy et al., 2018), and Maximum Class Probability (MCP) (Hendrycks & Gimpel, 2016). As shown in Table 4, DU demonstrates the best performance compared to other uncertainties. Among these uncertainties, entropy has a similar form to DU, however, it is unsuitable for our method, more details are given in Appendix B.2. Moreover, to validate the advantages of RC's relativity and asymmetric form, we conducted analysis and experiments in Appendix D.3.

6. Conclusion

Through extensive empirical studies, we observe that the fusion paradigms of existing methods are typically unreliable, and lack theoretical guarantees. Starting from the generalization error upper bound (GEB), we find the positive and negative correlations between fusion weight and loss, which inspired us to predict the Mono- and Holo-Confidence directly. Thus, we obtain predictable Co-Belief with theoretical guarantees to reduce GEB. Due to the potential prediction uncertainty, it is further calibrated in multimodal systems by relative calibration and used as the fusion weight. Comprehensive experiments with in-depth analysis validate our superiority in accuracy and stability against other approaches. Moreover, our PDF's extensions to other tasks are worth exploring. We believe this method is inspirational research that will benefit the community.

Impact Statement

This paper presents work to advance the field of multimodal fusion in machine learning. Our goal is to construct a predictive multimodal fusion method to boost the safety and accuracy of joint decisions in multimodal systems, lowering the potential modality bias and instability of prediction. However, due to the modality imbalance and data bias in open environments, there is a possibility of inevitable uncertainty when applying our method in real-world applications.

Acknowledgements

References

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14927–14937, 2020.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16:345–379, 2010.

Ayache, S., Quénot, G., and Gensel, J. Classifier fusion for svm-based multimedia semantic indexing. In *European Conference on Information Retrieval*, pp. 494–504. Springer, 2007.

Cao, B., Sun, Y., Zhu, P., and Hu, Q. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23555–23564, 2023.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Cui, H., Radosavljevic, V., Chou, F.-C., Lin, T.-H., Nguyen, T., Huang, T.-K., Schneider, J., and Djuric, N. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 2090–2096. IEEE, 2019.

Dempster, A. P. A generalization of bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232, 1968.

Denker, J. and LeCun, Y. Transforming neural-net output levels to probability distributions. In *Advances in Neural Information Processing Systems*, volume 3, 1990.

DeVries, T. and Taylor, G. W. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., and Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

Han, Z., Yang, F., Huang, J., Zhang, C., and Yao, J. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20707–20717, 2022a.

Han, Z., Zhang, C., Fu, H., and Zhou, J. T. Trusted multiview classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, 2022b.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, volume 34, pp. 677–689, 2021a.

Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., and Huang, L. What makes multi-modal learning better than single (provably). In *Advances in Neural Information Processing Systems*, volume 34, pp. 10944–10956, 2021b.

Huang, Z., Niu, G., Liu, X., Ding, W., Xiao, X., Wu, H., and Peng, X. Learning with noisy correspondence for cross-modal matching. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29406–29419, 2021c.

Kiela, D., Bhooshan, S., Firooz, H., Perez, E., and Testuggine, D. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Lee, J. and AlRegib, G. Gradients as a measure of uncertainty in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 2416–2420. IEEE, 2020.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Liu, W., Wang, X., Owens, J. D., and Li, Y. Energy-based out-of-distribution detection. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 21464–21475, 2020.

Liu, X., Zhu, X., Li, M., Wang, L., Tang, C., Yin, J., Shen, D., Wang, H., and Gao, W. Late fusion incomplete multiview clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2410–2423, 2018.

Ma, H., Han, Z., Zhang, C., Fu, H., Zhou, J. T., and Hu, Q. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6881–6893, 2021.

Ma, H., Zhang, Q., Zhang, C., Wu, B., Fu, H., Zhou, J. T., and Hu, Q. Calibrating multimodal learning. In *International Conference on Machine Learning*, pp. 23429–23450. PMLR, 2023.

Mackay, D. J. C. *Bayesian methods for adaptive models*. California Institute of Technology, 1992.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Natarajan, P., Wu, S., Vitaladevuni, S., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R., and Natarajan, P. Multimodal feature fusion for robust event detection in web videos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1298–1305. IEEE, 2012.

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Nefian, A. V., Liang, L., Pi, X., Liu, X., and Murphy, K. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2002:1–15, 2002.

Niu, T., Zhu, S., Pang, L., and El Saddik, A. Sentiment analysis on multi-view social data. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*, pp. 15–27. Springer, 2016.

Papadopoulos, G., Edwards, P. J., and Murray, A. F. Confidence estimation methods for neural networks: A practical comparison. *IEEE Transactions on Neural Networks*, 12(6):1278–1287, 2001.

Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. Balanced multimodal learning via on-the-fly gradient modulation.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022.

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.

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.

Scheunders, P. and De Backer, S. Wavelet denoising of multicomponent images using gaussian scale mixture models and a noise-free image as priors. *IEEE Transactions on Image Processing*, 16(7):1865–1872, 2007.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pp. 746–760. Springer, 2012.

Sim, T., Baker, S., and Bsat, M. The cmu pose, illumination, and expression database. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(12):1615 – 1618, December 2003.

Snoek, C. G., Worring, M., and Smeulders, A. W. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 399–402, 2005.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017.

Tempany, C. M., Jayender, J., Kapur, T., Bueno, R., Golby, A., Agar, N., and Jolesz, F. A. Multimodal imaging for improved diagnosis and treatment of cancers. *Cancer*, 121(6):817–827, 2015.

Wang, H., Yang, Y., and Liu, B. Gmc: Graph-based multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1116–1129, 2019a.

Wang, S., Liu, X., Zhu, E., Tang, C., Liu, J., Hu, J., Xia, J., and Yin, J. Multi-view clustering via late fusion alignment maximization. In *IJCAI*, pp. 3778–3784, 2019b.

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, X., Kumar, D., Thome, N., Cord, M., and Precioso, F. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6. IEEE, 2015.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., and Li, Y. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning*, pp. 23631–23644. PMLR, 2022.

Xie, Z., Wang, S. I., Li, J., Lévy, D., Nie, A., Jurafsky, D., and Ng, A. Y. Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573*, 2017.

Xue, Z. and Marculescu, R. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2574–2583, 2023.

Yan, R., Yang, J., and Hauptmann, A. G. Learning query-class dependent weights in automatic video retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 548–555, 2004.

Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

Zhang, Q., Wu, H., Zhang, C., Hu, Q., Fu, H., Zhou, J. T., and Peng, X. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pp. 41753–41769. PMLR, 2023.

Zhu, P., Sun, Y., Cao, B., and Hu, Q. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Appendix

A. Proofs

A.1. Proof of Theorem 3.1

Proof. Given the decision-level multimodal fusion formula delineated in Equation (1), consider ℓ to be the convex logistic loss function applied to binary classification tasks. The softmax function is utilized to normalize ω^m: $\omega^m = \frac{e^{\omega^m}}{\sum_{j=1}^{|\mathcal{M}|} e^{\omega^j}}$. Considering the property of convex function, we have:

$$\ell(f(x), y) = \ell(\sum_{m=1}^{|\mathcal{M}|} \omega^m f^m(x^{(m)}), y) \leq \sum_{m=1}^{|\mathcal{M}|} \omega^m \ell(f^m(x^{(m)}), y). \tag{12}$$

When computing the expectation of Equation (12) and leveraging the properties of expectation, the subsequent equation is satisfied. To simplify notation, $\ell(f^m(x), y)$ can be denoted as ℓ^m and $\mathcal{D}$ is an unknown dataset:

$$
\begin{aligned}
\mathrm{GE}(f) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(f(x), y) \\
&\leq \sum_{m=1}^{|\mathcal{M}|} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\omega^m \ell^m]) \\
&= \frac{1}{|\mathcal{M}|} \left(|\mathcal{M}| \cdot \sum_{m=1}^{|\mathcal{M}|} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\omega^m \ell^m] \right) \\
&= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|} \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}[\omega^m \ell^m] + (|\mathcal{M}| - 1)\mathbb{E}_{(x,y)\sim\mathcal{D}}[(1 - \sum_{j\neq m} \omega^j)\ell^m] \right) \\
&= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|} \Big(\mathbb{E}_{(x,y)\sim\mathcal{D}}[\omega^m] E_{(x,y)\sim\mathcal{D}}[\ell^m] + Cov(\omega^m, \ell^m) \\
&\quad - (|\mathcal{M}| - 1) \cdot \sum_{j\neq m} \big(\mathbb{E}_{(x,y)\sim\mathcal{D}}[\omega^j]\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell^m] + Cov(\omega^j, \ell^m) - \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell^m] \big) \Big) \\
&= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|} \Bigg[\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell^m] \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}[\omega^m] + (|\mathcal{M}| - 1) \cdot \Big[1 - \sum_{j\neq m} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\omega^j] \Big] \right) \\
&\quad + Cov(\omega^m, \ell^m) - (|\mathcal{M}| - 1) \sum_{j\neq m} Cov(\omega^j, \ell^m) \Bigg] \\
&= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|} \Bigg[\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell^m] \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}[\omega^m] + (|\mathcal{M}| - 1) \cdot \mathbb{E}_{(x,y)\sim\mathcal{D}}[\omega^m] \right) \\
&\quad + Cov(\omega^m, \ell^m) - (|\mathcal{M}| - 1) \sum_{j\neq m} Cov(\omega^j, \ell^m) \Bigg] \\
&= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|} \Bigg[|\mathcal{M}| \cdot \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell^m]\mathbb{E}_{(x,y)\sim\mathcal{D}}[\omega^m] + Cov(\omega^m, \ell^m) - (|\mathcal{M}| - 1) \sum_{j\neq m} Cov(\omega^j, \ell^m) \Bigg] \\
&= \sum_{m=1}^{|\mathcal{M}|} \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell^m]\mathbb{E}_{(x,y)\sim\mathcal{D}}[\omega^m] + \frac{1}{|\mathcal{M}|} \cdot \Big[Cov(\omega^m, \ell^m) - (|\mathcal{M}| - 1) \sum_{j\neq m} Cov(\omega^m, \ell^j) \Big] \right) \\
&\leq \sum_{m=1}^{|\mathcal{M}|} \left(\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell^m] + \frac{1}{|\mathcal{M}|} Cov(\omega^m, \ell^m) - \frac{|\mathcal{M}| - 1}{|\mathcal{M}|} \sum_{j\neq m} Cov(\omega^m, \ell^j) \right)
\end{aligned}
$$

To simplify Equation (13), we invoke Rademacher complexity theory (Mohri et al., 2018) (Theorem 3.5), which establishes that with a confidence level of $1 - \Delta$ where $0 < \Delta < 1$, the following holds:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell^m] \leq e\hat{r}r[f^m] + \mathcal{R}_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\Delta)}{2N}}. \tag{13}$$

In this context, $e\hat{r}r(f^m)$ represents the empirical error of the unimodal function f^m, and $\mathcal{H}$ denotes the hypothesis set, defined as $\mathcal{H} : \mathcal{X} \to \{-1, +1\}$, which includes f as a member. The Rademacher complexity is denoted by $\mathcal{R}_N(\mathcal{H})$. Consequently, we assert that with a confidence level of $1 - \Delta$, where $0 < \Delta < 1$, the following relationship is upheld:

$$\text{GE}(f) \leq |\mathcal{M}| \left(\mathcal{R}_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\Delta)}{2N}} \right) + \sum_{m=1}^{|\mathcal{M}|} e\hat{r}r(f^m) + \sum_{m=1}^{|\mathcal{M}|} \left[\frac{1}{|\mathcal{M}|} \underbrace{Cov(\omega^m, \ell^m)}_{\text{Mono-Covariance}} - \frac{|\mathcal{M}|-1}{|\mathcal{M}|} \sum_{j\neq m} \underbrace{Cov(\omega^m, \ell^j)}_{\text{Holo-Covariance}} \right]. \tag{14}$$

A.2. Proof of Mono-Confidence Satisfies Corollary 3.2

Proof. In classification tasks with a cross-entropy function, the unimodal loss is defined as:

$$\ell = -\sum_{i=1}^{N} y_i \log p_i. \tag{15}$$

The one-hot label of the i-th class is denoted by y_i, and p_i represents the predicted probability for the i-th class. Under the assumption that the t-th class is the correct classification, we have $y_t = 1$ and $y_i = 0$ for all $i \neq t$. The probability of true class p_{true} is p_t under the above assumption. Consequently, Equation (15) can be simplified to:

$$\ell = -y_t \log p_t - \sum_{i \neq t} y_i \log p_i$$

$$= -\log p_t \tag{16}$$

$$= -\log p_{true} \tag{17}$$

Noting that p_{true} naturally correlates with cross-entropy loss. To substantiate that $Cov(p_{true}^m, \ell^m) < 0$, we put forth the following proposition:

Proposition A.1. *For any two random variables X and Y, the condition $Cov(X, Y) < 0$ is equivalent to that X and Y are inversely correlated, and conversely, $Cov(X, Y) > 0$ is equivalent to a positive correlation between them.*

Recalling Equation (17), we have

$$\frac{\mathrm{d}\ell}{\mathrm{d}p_{true}} = -\frac{1}{p_{true}}, \tag{18}$$

where $p_{true} \in [0, 1]$. Hence, we have $\frac{\mathrm{d}\ell}{\mathrm{d}p_{true}} \in (-\infty, -1]$, which means ℓ is negatively correlated with p_{true} in the domain of p_{true}. With the Proposition A.1, the fact that $Cov(p_{true}^m, \ell^m) < 0$ holds. Recalling the Corollary 3.2, it's proved that using p_{true} as the Mono-Confidence conforms to reducing generalization error upper bound.

A.3. Proof of Holo-Confidence Satisfies Corollary 3.3

Proof. Recalling the definition of the m-th modality's Holo-Confidence:

$$\text{Holo-Conf}^m = \frac{\sum_{j\neq m} \ell^j}{\sum_{i=1}^{|\mathcal{M}|} \ell^i}. \tag{19}$$

As for the positive correlation between Holo-Conf^m and ℓ^j, the derivative of Holo-Conf^m with respect to ℓ^j for all $j \neq m$ is computed as follows:

$$\frac{\partial \text{Holo-Conf}^m}{\partial \ell^j} = \frac{\sum_{i=1}^{|\mathcal{M}|} \ell^i - \sum_{j\neq m} \ell^j}{\left(\sum_{i=1}^{|\mathcal{M}|} \ell^i \right)^2} = \frac{\ell^j}{\left(\sum_{i=1}^{|\mathcal{M}|} \ell^i \right)^2}, \tag{20}$$

where $\ell \in [0, +\infty)$. Consequently, since $\frac{\mathrm{dHolo\text{-}Conf}^m}{\mathrm{d}\ell^j} \in [0, +\infty)$, it is established that Holo-Conf^m is positively correlated with ℓ^j within the domain of ℓ^j, for all $j \neq m$. Given Proposition A.1, it is validated that $Cov(\text{Holo-Conf}^m, \ell^j) > 0$ for all $j \neq m$. Referencing Corollary 3.3, this demonstrates that our proposed Holo-Confidence metric aligns with the reduction of the generalization error upper bound.

A.4. Proof of Co-Belief Lowers GEB

To verify whether a fusion strategy can reduce the model's GEB, we proposed the **A**ggregate **C**ovariance (AC), where

$$\text{AC}(f) = \sum_{m=1}^{|\mathcal{M}|} \left(Cov(\omega^m, \ell^m) - (|\mathcal{M}| - 1) \sum_{j \neq m} Cov(\omega^m, \ell^j) \right), \tag{21}$$

where $f \in \mathcal{H}$ is the multimodal function and f^m is the function of m-th modality. When $\text{AC}(f) < 0$, it is deemed that the GEB of f is decreased. We identify that reducing the generalization upper bound requires ensuring our proposed metric, $\text{AC}(f) < 0$. The formulation of the proposed Co-Belief is as follows:

$$\text{Co-Belief}^m = p_{true}^m + \frac{\log \prod_{j \neq m} p_{true}^j}{\log \prod_{i=1}^{|\mathcal{M}|} p_{true}^i} = e^{-\ell^m} + \frac{\sum_{j \neq m} \ell^j}{\sum_{i=1}^{|\mathcal{M}|} \ell^i}. \tag{22}$$

The desirable result is:

$$\text{AC}(f) = \sum_{m=1}^{|\mathcal{M}|} \left(\underbrace{Cov(\text{Co-Belief}^m, \ell^m)}_{\text{Mono-Covariance}} - (|\mathcal{M}| - 1) \sum_{j \neq m} \underbrace{Cov(\text{Co-Belief}^m, \ell^j)}_{\text{Holo-Covariance}} \right) < 0. \tag{23}$$

Now, we consider the Mono-Covariance in Equation (23).

$$\frac{\mathrm{dCo\text{-}Belief}^m}{\mathrm{d}\ell^m} = -e^{\ell^m} - \frac{\sum_{j \neq m} \ell^j}{\left(\sum_{i=1}^{|\mathcal{M}|} \ell^i\right)^2}, \tag{24}$$

where $\ell \in [0, +\infty)$. Hence, we have $\frac{\mathrm{dCo\text{-}Belief}^m}{\mathrm{d}\ell^m} \in (-\infty, -1]$, which means Co-Belief^m is negatively correlated with ℓ^m in the domain of ℓ^m. Recalling Proposition A.1, we have Mono-Covariance < 0.

Then, consider the Holo-Covariance in Equation (23).

$$\frac{\mathrm{dCo\text{-}Belief}^m}{\mathrm{d}\ell^j} = \frac{\sum_{i=1}^{|\mathcal{M}|\ell^i - \sum_{j \neq m} \ell^j}}{\left(\sum_{i=1}^{|\mathcal{M}|} \ell^i\right)^2} = \frac{\ell^j}{\left(\sum_{i=1}^{|\mathcal{M}|} \ell^i\right)^2}, \tag{25}$$

where $\ell \in [0, \infty)$. Therefore, $\frac{\mathrm{dCo\text{-}Belief}^m}{\mathrm{d}\ell^j} \in [0, +\infty)$ indicates that Co-Belief^m is positively correlated with ℓ^j in the domain of $\ell^j, \forall j \neq m$. With the Proposition A.1, we have Holo-Covariance > 0. Thus, we achieve our goal in Equation (23), reducing the generalization error upper bound.

Furthermore, our proposed Co-Belief surpasses both Mono-Confidence and Holo-Confidence in reducing the generalization error upper bound. Unlike Mono-Conf^m, which exhibits no correlation with ℓ^j, the Mono-Covariance of Co-Belief^m is less than that of Mono-Conf^m. Similarly, Holo-Conf^m has a higher Mono-Covariance than Co-Belief^m. These findings underscore our proposed Co-Belief's marked advantage in diminishing the generalization error upper bound, as corroborated by our ablation studies.

A.5. The Complete Form of the RC Formula

Revisiting our proposed relative calibration term as defined in Equation (8), it is important to note that this term is initially conceptualized under a two-modality setting. For cases where $|\mathcal{M}| > 2$, the relative calibration term is redefined as follows:

$$\text{RC}^m = \frac{\text{DU}^m \cdot (|\mathcal{M}| - 1)}{\sum_{i \neq m} \text{DU}^i}. \tag{26}$$

Moreover, when formulating the final calibrated Co-Belief, we perform truncation on the RC value for asymmetric calibration, which is defined by the following formula:

$$\mathrm{k}^m = \begin{cases} \mathrm{RC}^m = \frac{\mathrm{DU}^m \cdot (|\mathcal{M}| - 1)}{\sum_{i \neq m} \mathrm{DU}^i} & \text{if } \mathrm{RC}^m < 1, \\ 1 & \text{otherwise.} \end{cases} \tag{27}$$

It is observed that when $|\mathcal{M}| = 2$, the definition is congruent with the formulation of RC and k as delineated in Equation (8) and Equation (9). Furthermore, this definition is consistent with the theoretical analysis presented in the main body of our paper.

B. More Analysis

B.1. $\hat{p}_{true}$ is Reliable

Existing confidence estimation methods (Corbière et al., 2019; Papadopoulos et al., 2001) usually depend on estimating confidence precisely for certain tasks in unimodal scenarios. By contrast, in our proposed method, p_{true} is used to construct dynamic fusion weight to satisfy the theoretical guarantee and reflect the modality dominance in multimodal fusion. Therefore, the reliability of $\hat{p}_{true}$ in our method is reflected in the ability to provide reasonable fusion weights (Mono-Confidence) for each modality, which implies that when different modalities make inconsistent decisions, the fusion weight of the dominant modality should be higher and vice versa, to make a correct decision jointly.

To validate the reliability of $\hat{p}_{true}$, We calculated the probability that $\hat{p}_{true}$ is able to help make the right decision in case unimodal models make false classifications. For the two-modalities dataset, we first counted the number of samples in which the two modalities made conflicting decisions. Among these conflicting samples, we then calculated the proportion of cases where the $\hat{p}_{true}$-weighted (Mono-Confidence weighted) fusion result was correct. We compare $\hat{p}_{true}$-weighted fusion with other methods that have uni-modal outputs. As shown in Table 5, the experimental results indicate that $\hat{p}_{true}$ is superior in responding to the importance of modality and correct for the modality results' inconsistencies compared with other fusion weights, demonstrating the reliability of our $\hat{p}^m_{true}$.

B.2. Drawbacks of Employing Entropy in Composing Relative Calibration

Traditional approaches to evaluating informational uncertainty, exemplified by entropy (Shannon, 1948), face difficulties within our relative framework. Figure 6 (a) demonstrates that when entropy is held constant for a particular modality, the variable rate of change of its derivative engenders erratic fluctuations in the relative value of multimodal entropy in response to variations in another modality's entropy, resulting in biased comparisons. In contrast, Distribution Uniformity (DU), characterized by its constant slope, facilitates more balanced assessments of uncertainty across different modalities.
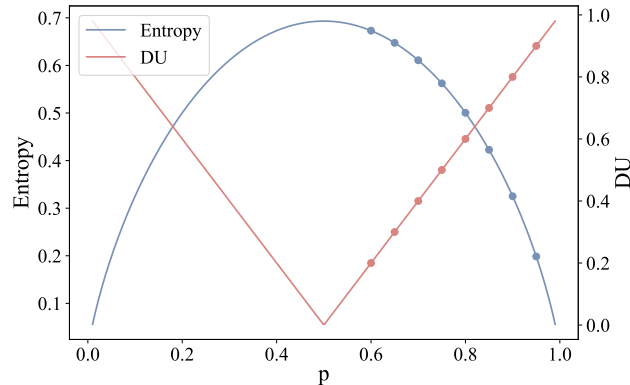

Figure 6. The function diagram of Entropy and Distribution Uniformity in two categories.

Table 5. We report the proportion (%) of fusion results that are correct by different fusion weight when the two modalities made conflicting decision.

METHOD	MVSA	UMPC FOOD 101
LATE FUSION	60.58	89.31
TMC	60.09	90.84
QMF	73.45	93.77
DYNMM	76.77	88.94
OURS	**77.49**	**95.53**

B.3. More Comparisons with Other Methods

The recognition of decision-level generalization bounds in past analyses by QMF is commendable. Nevertheless, it is crucial to acknowledge that our methodology diverges markedly at the conceptual stage from QMF's approach. QMF's formula for generalization bounds, grounded in unimodal analysis, overlooks the complex interplay among modalities within a multimodal system. The terms 'Term-L' and 'Term-C' within QMF's framework are not static, varying with each modality's response to scene changes. Consequently, enforcing a negative Term-Cov through calibration does not ensure a reduced generalization bound. Our perspective posits that each modality's contribution in a multimodal system should be assessed in relation, and with this view, we have deduced the Generalization Error Bound (GEB) through a multimodal perspective. Fundamentally, the GEB is influenced only by Mono-Term and Holo-Term, considering that optimization within the same function class yields identical Rademacher complexity and empirical error. Therefore, ensuring the sign of Mono-Term and Holo-Term suffices to diminish the generalization bound.

Furthermore, QMF endeavors to gauge the quality of modalities by accounting for uncertainty, postulating a direct relationship between uncertainty and loss, thereby aligning its approach with its theoretical framework of the generalization bound. This necessitates the application of calibration to uphold the initial premise. However, our analysis of the p_{true}, which indicates the modality's reliability, reveals an inverse relationship with the loss. Consequently, rooted in the Generalization Error Bound (GEB) concept, we propose a strategy to diminish the GEB by focusing on the accurate prediction of p_{true}.

C. More Details

C.1. Related Work Details

Predictive confidence is currently frequently used in fault detection (Corbière et al., 2019) and Out-Of-Distribution (OOD) detection (DeVries & Taylor, 2018). However, it often leads to overconfidence due to softmax probabilities. Some methods focus on smoothing the predicted probability distribution through label smoothing (Müller et al., 2019), while others apply temperature scaling (Liang et al., 2017) to calibrate the probability outputs. Some works (Wei et al., 2022) mitigate overconfidence by constraining the magnitude of logits. Essentially, most of these approaches aim to align the expected class probabilities with empirical accuracy (Ma et al., 2023).

C.2. Symbols Table

To avoid potential confusion, we provide a table for main symbols in Table 6.

C.3. Metric Details

To quantify the capacity of a fusion strategy on reducing the GEB, we defined a metric called **GEB Decreasing Proportion** (GDP):

$$\text{GDP} = \mathbb{E}_{\mathcal{F}}[\mathbb{I}_{\{\text{AC}(f)<0|f\in\mathcal{H}\}}(f)], \tag{28}$$

where $\mathcal{F} \subset \mathcal{H}$, $\mathbb{I}$ is the indicator function, which is defined as:

$$\mathbb{I}_{\{\text{AC}(f)<0|f\in\mathcal{H}\}}(f) = \begin{cases} 1 & \text{if } f \in \{\text{AC}(f) < 0|f \in \mathcal{H}\}, \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

Table 6. Main Symbols Table.

SYMBOL	EXPLANATION
$\mathcal{M}$	THE SET OF EACH UNI-MODAL
$|\mathcal{M}|$	THE CARDINALITY OF $\mathcal{M}$
ω^m	FUSION WEIGHT OF THE M-TH MODALITY
f^m	UNI-MODAL PROJECTION FUNCTION
$\text{GEB}(f)$	GENERALIZATION ERROR UPPER BOUND OF f
p_{true}	TRUE CLASS PROBABILITY
$\hat{p}_{true}$	THE PREDICTION OF p_{true}
ℓ	LOGISTIC LOSS FUNCTION
ℓ^m	THE SIMPLICITY REPRESENTATION OF $\ell(f^m(x^m), y)$
$\hat{\ell^m}$	THE PREDICTION OF ℓ^m
$\hat{err}(f^m)$	EMPIRICAL ERRORS OF THE M-TH MODALITY
$\mathcal{H}$	HYPOTHESIS SET
$R_N(\mathcal{H})$	RADEMACHER COMPLEXITIES

C.4. Implementation Details

The network was trained for 100 epochs utilizing the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 0.01, dropout rate of 0.1, and a batch size of 16. The initial learning rate was chosen from the set $\{1e-8, 5e-5, 1e-4\}$. Specifically, for image-text classification, the initial learning rate was $5e-5$; for scene recognition, it was 1e-8 for the second layer of the confidence predictor and 1e-4 for all others; for emotion recognition, it was set to 1e-3. All the experiments were conducted on an NVIDIA A6000 GPU, using PyTorch with default parameters for all methods.

C.5. Experiment Details

We changed the data quality by varying noise levels. Specifically, we alter the degree of noise added to it and fix another modality's noise level. For NYU Depth V2 dataset, we fixed the RGB noise level at 5 and increased the depth noise from 0 to 10. For MVSA and FOOD101 datasets, we maintained the text noise at 2.5 and escalated the image noise from 0 to 5. For the CREMA-D dataset, we kept the audio modality's SNR fixed and varied the image noise from 0 to 10. With the increase of one modality's noise, we report the changing trend of each modality's calibrated weight.

C.6. The Prediction of p_{true}^m and Loss Function

During the inference phase, there is no ground truth available to get the p_{true}^m, so we trained a confidence predictor consisting of multiple linear layers to predict $\hat{p}_{true}^m$ by the MSELoss:

$$\mathcal{L}_{p_{true}} = \sum_{m=1}^{|\mathcal{M}|} MSE\left(\hat{p}_{true}^m, p_{true}^m\right), \tag{30}$$

where $\hat{p}_{true}^m = \text{Predictor}(feature^m)$, and $feature^m$ is the feature of input x^m generated by encoder. Leveraging the $\hat{p}_{true}^m$ to compute Co-Belief^m, utilizing the model's output after softmax to calculate RC^m and its transformation k_m, the final calibrated Co-Belief CCB^m as well as ω^m can be obtained as Equation (10).

Drawing on the principles of multi-task learning, we conceive the overall loss function as the aggregate of standard cross-entropy classification losses across multiple modalities, coupled with the p_{true} prediction loss:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{CE}}(y, f(x)) + \sum_{m=1}^{|\mathcal{M}|} \mathcal{L}_{\text{CE}}(y, f^m(x^m)) + \mathcal{L}_{p_{true}}, \tag{31}$$

where $\mathcal{L}_{\text{CE}}$ represents the cross-entropy loss, while $\mathcal{L}_{p_{true}}$ is the p_{true} prediction loss defined as Equation (30).
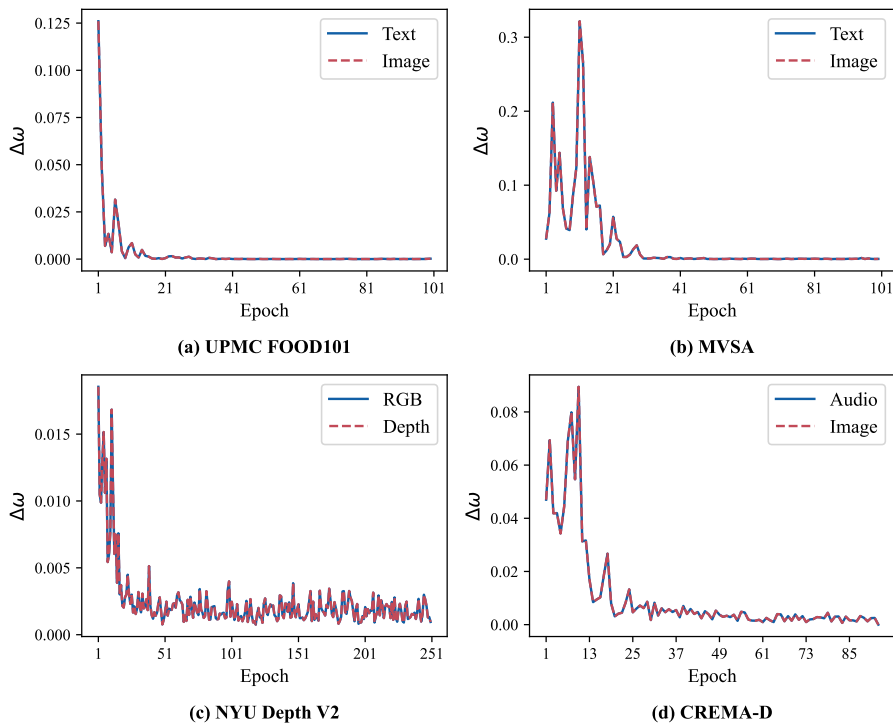
Figure 7. $\Delta\omega^m$ gradually shrinks to zero as the training epochs increase.

Table 7. Comparative experiments on asymmetric calibration and its variants with NYU Depth V2.

VARIANTS OF RC	$\epsilon = 0$	$\epsilon = 5$	$\epsilon = 10$
DISTRIBUTION UNIFORMITY (DU)	71.13	64.93	62.08
RELATIVE CALIBRATION (RC)	71.06	65.43	62.19
ASYMMETRIC CALIBRATION	**71.37**	**65.72**	**62.56**

D. Additional results

D.1. Lower Generalization Error Upper Bound

We applied the proposed Generalization Error Bound Decreasing Proportion (GDP, as defined in Equation (28)) to measure the capacity of a fusion strategy for reducing the Generalization Error Upper Bound (GEB). Separately, we compare the GDP metrics of models that utilize Mono-Conf (Equation (4)), Co-Belief (Equation (5)), and CCB (Equation (10)) as fusion weights. Specifically, for each fusion strategy, we trained 50 models with distinct random seeds and reported the GDP of these models under varying noises (0, 5, and 10) during testing. We took Mono-Confidence, Co-Belief, and CCB as distinct fusion weights for the model and depicted the AC distribution in Figure 3 (a), (b), and (c). The proportion of AC < 0, i.e. GDP, signifies the fusion strategy's potential to reduce the model's GEB. Figure 3 (d) displays the GDP of different fusion strategies under diverse noisy conditions. The proportion of Mono-Confidence that reduces GEB without noise is greater than 50%, and Holo-Confidence further increases the model's GDP, CCB endows the model with the best generalization capabilities in dynamic environments. These experiments validate our effectiveness in lowering the generalization error upper bound.

On comparing Figure 3 (a) and (b), it is evident that the extent of the orange segment in Figure 3 (b), i.e., the model's GDP, is greater than that in Figure 3 (a), suggesting that the integration of both Mono-Confidence and Holo-Confidence yields a heightened likelihood of GEB reduction. In Figure 3 (c), the application of Calibrated Co-Belief as the fusion weight demonstrates a probability of up to 90% in reducing the generalization error.

Table 8. Comparison on three modalities dataset PIE under Gaussian noise.

METHOD	$\epsilon = 0.0$	$\epsilon = 2.0$	$\epsilon = 4.0$	$\epsilon = 6.0$	$\epsilon = 8.0$	$\epsilon = 10.0$
LATE FUSION	88.24	85.74	83.09	81.32	78.09	74.12
TMC	89.71	84.26	79.12	74.71	70.88	64.85
QMF	88.24	85.59	81.76	81.03	77.94	74.00
DYNMM	89.71	85.29	84.56	82.35	80.15	76.47
OURS	**90.44**	**88.82**	**84.71**	**82.50**	**80.74**	**78.53**

Figure 3 (d) displays the GDP of different fusion strategies under various noisy conditions, with the sum corresponding to the GDP of the three fusion strategies depicted in Figure 3 (a) (b) and (c). We observe that the proportion of Mono-Confidence reducing GEB without noise is greater than 50%, and the inclusion of Holo-Confidence further increases the model's GDP, indicating an enhanced generalization ability. Additionally, the incorporation of the calibration strategy endows the model with stronger generalization capabilities in noisy environments.

D.2. Convergence of Fusion Weight

To demonstrate the dependability of our calibrated weights, we ascertain their convergence when training. We define $\Delta\omega^m$ as the mean absolute change of the weights across the entire validation set for each epoch and track its progression throughout the training period. As illustrated in Figure 7, across various datasets, $\Delta\omega^m$ consistently trends toward zero, indicating the calibrated fusion weights' convergence.

D.3. Effectiveness of the Form of the Asymmetric Calibration

We adopt an asymmetric form for Relative Calibration (RC), which not only aligns with the motivation of calibration but also facilitates better weight optimization for the RC with an asymmetric form ranging from 0 and 1. To validate the effectiveness of asymmetric calibration, we conducted comparative experiments involving asymmetric calibration and its variants, including DU (Equation (7)), and RC (Equation (8)) without asymmetric form by assessing their impact on average accuracy metrics. Results presented in Table 7 suggest that RC with asymmetric calibration exhibits enhanced generalization capabilities.

D.4. Extensibility to Datasets with More than 2 Modalities

As illustrated in Equation (27), our proposed relative calibration can be extended to cases where $|\mathcal{M}| > 2$. To verify the effectiveness of our PDF, we conducted comparisons with previous state-of-the-art methods on the PIE dataset with three modalities. As depicted in Table 8, our PDF surpasses the competing methods across various noise levels, highlighting its superiority.

D.5. Compared Experiments on Different Noises

In this section, we report the full experiment results with standard deviation in varying Gaussian noise and Salt-pepper noise compared with other methods in Table 9 and Table 10 separately.

E. Limitations

Even though the proposed PDF model achieves superior performance over existing methods and shows advanced generalization ability in dynamically changing conditions, there are still some potential limitations. We provide theoretical guarantees to the Co-Belief prediction, however, the potential uncertainty is inevitable. The proposed relative calibration is an empirical solution for this problem without theoretical guarantees. Therefore, it is important to explore new uncertainty estimation methods from the theoretical perspective. Besides, the predictor in our model is relatively simple, and it is also valuable to study more efficient and effective predictor architectures.

Table 9. We add Gaussian noise on 50% modalities and ϵ presents the noise degree. it shows the average and the standard deviation of classification accuracies with our method and the compared methods on four datasets. The method marked with * was replicated by us, while the rest of the data is sourced from (Zhang et al., 2023).

DATASET	METHOD	$\epsilon = 0.0$	$\epsilon = 5.0$	$\epsilon = 10.0$
MVSA	TEXT	75.61±0.53	69.50±1.50	47.41±0.79
	IMG	64.12±1.23	49.36±2.02	45.00±2.63
	CONCAT	65.59±1.33	50.70±2.65	46.12±2.44
	LATE FUSION	76.88±1.30	63.46±3.46	55.16±3.60
	QMF	78.07±1.10	73.85±1.42	61.28±2.12
	TMC	74.87±2.24	66.72±4.55	60.35±2.79
	DYNMM*	79.07±0.53	67.96±1.65	59.21±1.41
	OURS	**79.94±0.95**	**74.4±1.51**	**63.09±1.33**
UMPC FOOD 101	TEXT	86.46±0.05	67.38±0.19	43.88±0.32
	IMG	64.62±0.40	34.72±0.53	33.03±0.37
	CONCAT	88.20±0.34	61.10±2.02	49.86±2.05
	LATE FUSION	90.69±0.12	68.49±3.37	57.99±1.59
	QMF	92.92±0.11	76.03±0.70	62.21±0.25
	TMC	89.86±0.07	73.93±0.34	61.37±0.21
	DYNMM*	92.59±0.07	74.74±0.19	59.68±0.20
	OURS	**93.32±0.22**	**76.47±0.31**	**62.83±0.31**
NYU DEPTH V2	RGB	62.65±1.22	50.95±3.38	44.13±3.80
	DEPTH	63.30±0.48	53.12±1.52	45.46±2.07
	CONCAT*	69.88±0.52	63.82±1.46	60.03±2.63
	LATE FUSION*	70.03±0.84	64.37±0.80	60.55±1.65
	TMC*	70.40±0.31	59.33±2.19	50.61±2.87
	QMF*	69.54±1.06	64.10±1.42	60.18±1.23
	DYNMM*	65.50±0.37	54.31±1.72	46.79±1.09
	OURS	**71.37±0.76**	**65.72±1.72**	**62.56±1.84**
CREMA-D	VISUAL*	43.60±2.02	32.52±1.98	30.17±1.19
	AUDIO*	58.67±1.01	54.66±2.16	43.01±6.04
	CONCAT*	61.56±1.37	52.33±3.32	41.01±5.70
	LATE FUSION*	61.81±2.13	49.84±3.72	39.15±5.82
	QMF*	63.04±1.37	56.60±2.38	41.60±2.75
	TMC*	59.15±1.95	54.42±3.34	46.79±4.72
	DYNMM*	60.46±0.37	54.43±0.73	42.39±0.50
	OURS	**63.31±1.11**	**57.85±2.04**	**47.84±2.32**

Table 10. We add Salt-pepper noise on 50% modalities and ϵ presents the noise degree. it shows the average and the standard deviation of classification accuracies with our method and the compared methods on four datasets. The method marked with * was replicated by us, while the rest of the data is sourced from (Zhang et al., 2023).

DATASET	METHOD	$\epsilon = 0.0$	$\epsilon = 5.0$	$\epsilon = 10.0$
MVSA	TEXT	75.61±0.53	69.50±1.50	47.41±0.79
	IMG	64.12±1.23	56.72±1.92	50.71±3.20
	CONCAT	65.59±1.33	58.69±2.25	51.16±2.99
	LATE FUSION	76.88±1.30	67.88±1.87	55.43±1.94
	QMF	78.07±1.10	73.90±1.89	60.41±2.63
	TMC	74.87±2.24	68.02±3.07	56.62±3.67
	DYNMM*	79.07±0.53	71.35±0.97	59.96±1.31
	OURS	**79.94±0.95**	**75.11±1.15**	**61.97±1.14**
UMPC FOOD101	TEXT	86.44±0.02	67.41±0.20	43.89±0.33
	IMG	64.53±0.47	50.75±0.44	36.83±0.92
	CONCAT	88.22±0.36	72.49±0.75	52.10±0.97
	LATE FUSION	90.66±0.16	77.99±0.54	58.75±0.99
	QMF	92.90±0.13	80.87±0.40	61.60±0.20
	TMC	89.86±0.07	77.86±0.41	60.22±0.43
	DYNMM*	92.59±0.07	78.91±0.20	57.64±0.30
	OURS	**93.32±0.22**	**81.21±0.34**	**61.76±0.33**
NYU DEPTH V2	RGB	62.61±1.21	49.14±1.40	34.76±1.59
	DEPTH	63.32±0.50	50.99±1.41	38.56±2.16
	CONCAT*	69.88±0.52	61.41±1.69	51.65±2.94
	LATE FUSION*	70.03±0.84	62.05±1.17	51.50±1.81
	TMC*	70.40±0.31	59.33±1.47	45.32±2.84
	QMF*	69.54±1.06	62.02±1.47	51.87±0.91
	DYNMM*	65.50±0.37	52.26±1.45	38.17±1.17
	OURS	**71.37±0.76**	**64.27±1.36**	**53.62±2.15**
CREMAD	VISUAL*	43.60±2.02	40.30±1.77	36.84±1.72
	AUDIO*	58.67±1.01	54.57±2.06	43.00±6.01
	CONCAT*	61.56±1.37	54.28±3.89	42.57±6.16
	LATE FUSION*	61.81±2.13	54.83±3.24	41.07±6.88
	QMF*	63.04±1.37	57.73±2.25	45.02±2.28
	TMC*	59.15±1.95	54.61±3.19	47.72±2.76
	DYNMM*	60.46±0.37	54.58±0.65	42.49±0.43
	OURS	**63.31±1.11**	**58.61±1.50**	**48.40±2.85**