Unbiased Dynamic Multimodal Fusion

Shicai Wei, Kaijie Zhang, Luyi Chen, Tao He, Guiduo Duan*
University of Electronic Science and Technology of China
shicaiwei@uestc.edu.cn,duanguiduo@163.com

Abstract

Traditional multimodal methods often assume static modality quality, which limits their adaptability in dynamic real-world scenarios. Thus, dynamical multimodal methods are proposed to assess modality quality and adjust their contribution accordingly. However, they typically rely on empirical metrics, failing to measure the modality quality when noise levels are extremely low or high. Moreover, existing methods usually assume that the initial contribution of each modality is the same, neglecting the intrinsic modality dependency bias. As a result, the modality hard to learn would be doubly penalized, and the performance of dynamical fusion could be inferior to that of static fusion. To address these challenges, we propose the Unbiased Dynamic Multimodal Learning (UDML) framework. Specifically, we introduce a noise-aware uncertainty estimator that adds controlled noise to the modality data and predicts its intensity from the modality feature. This forces the model to learn a clear correspondence between feature corruption and noise level, allowing accurate uncertainty measure across both low- and high-noise conditions. Furthermore, we quantify the inherent modality reliance bias within multimodal networks via modality dropout and incorporate it into the weighting mechanism. This eliminates the dual suppression effect on the hard-to-learn modality. Extensive experiments across diverse multimodal benchmark tasks validate the effectiveness, versatility, and generalizability of the proposed UDML. The code is available at https://github.com/shicaiwei123/UDML.

1. Introduction

Multimodal learning has achieved significant progress across a wide range of vision tasks, including classification [12, 22, 39], object detection [18, 30, 52], and segmentation [3, 14, 28]. Most existing multimodal learning methods assume that the quality of each modality is static, with one modality consistently stronger than the others. However, this assumption often fails in real-world scenarios. For example, the RGB modality typically provides richer information than the IR modality during daytime but suffers substantial reliability degradation at night, whereas the IR modality often outperforms RGB under low-light conditions [20]. Therefore, it is essential to develop multimodal learning methods capable of handling dynamic data quality.

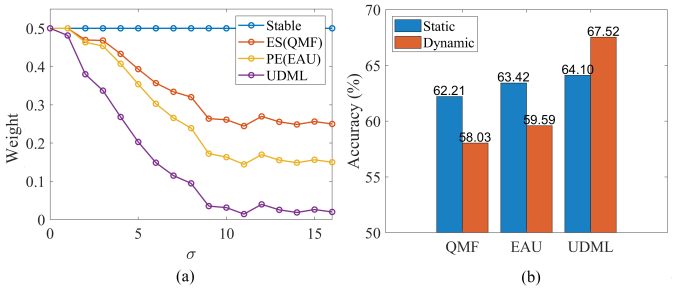

Figure 1. Visualization of dynamic multimodal methods for audio-visual classification on the CREMA-D dataset. (a) Visual weighting coefficients obtained using different uncertainty estimation methods, such as energy score (ES) [47] and probabilistic embedding (PE) [8], and the proposed UDML, as varying levels of noise (σ) are injected into the visual modality. (b) Performance Comparison of different methods under static and dynamic weighting when noise ($\sigma = 5$) is injected into the visual modality.

To address the dynamic quality of multimodal data, numerous methods have been proposed to assess modality reliability and adjust their contributions accordingly. These methods can be broadly categorized into two paradigms: prior-based methods [9, 20] and uncertainty-based methods [8, 11, 16, 47]. Prior-based methods estimate modality quality using human knowledge or experience. For instance, LiDAR sensors may be prioritized at night while cameras are favored in daylight in autonomous driving. While intuitive, such approaches often lack generalizability, particularly in complex or unforeseen environments. Thus, uncertainty-based methods have recently gained attention as a more principled and generalizable solution. By leveraging probabilistic modeling or information-theoretic measures, these methods dynamically adjust modality contributions, offering a more robust foundation for reliable multimodal integration.

Despite advances in dynamic multimodal learning, exist-

ing approaches still suffer from two major limitations. First, most existing uncertainty estimation methods rely on empirical metrics, such as energy score [47] and probabilistic embedding [8, 16], failing to assess the modality quality when noise levels are either too low or too high. Taking the widely used probabilistic embedding (PE) as an example (Fig. 1(a)), we observe that PE fails to detect low-intensity noise($\sigma < 4$), leading to rigid modality weighting. Besides, even under extreme corruption ($\sigma > 10$), PE still assigns considerable weight to the corrupted modality rather than disregarding it. Second, existing methods generally assume that the initial contributions of all modalities are the same, overlooking modality dependence biases. In practice, multimodal models tend to rely more on easy-to-learn modalities to make decisions [17, 27]. As a result, the modality hard to learn would be doubly penalized due to optimization bias and high uncertainty. Consequently, as shown in Fig 1(b), the performance of dynamic fusion could be inferior to that of the stable fusion.

To this end, we propose an Unbiased Dynamic Multimodal Learning (UDML), a general framework to assist multimodal learning with dynamical data quality. UDML consists of two key components: a noise-aware uncertainty estimator and a modality-dependency calculator. The noise-aware uncertainty estimator adds controlled noise to the modality data and predicts its intensity from the modality feature, achieving accurate estimation across both low- and high-noise regimes. To further ensure robustness to unseen noise types, we introduce the probabilistic representation technique that maps each modality into a distribution to decouple noise from semantic content. Specifically, the mean encodes semantic information, and the variance reflects noise characteristics. The estimator then derives noise intensity directly from the variance. The modality-dependency calculator employs modality dropout to quantify the output's inherent reliance on each modality. This dependency measure is then used to recalibrate modality weights, balancing their contributions and mitigating dual-suppression effects. Notably, UDML is architecture-agnostic, as all components operate solely on modality representations, ensuring generalizability across diverse models and fusion methods. Finally, we introduce a progressive optimization strategy that enables the simultaneous learning of multimodal representations, noise estimation, and the primary task within a standard training schedule. Extensive experiments on multiple tasks and datasets verify that UDML consistently enhances multimodal performance, demonstrating its robustness and versatility.

- We reveal the bias of existing uncertainty-based estimators, which are insensitive to slight degradation and assign non-negligible weights to heavily corrupted modalities. This limits the robustness of dynamic fusion.
- We reveal the dual suppression effect in existing dynamic multimodal learning, in which the modality hard to learn is doubly penalized due to optimization bias and high uncertainty. This could lead to the dynamic fusion underperforming the static fusion.
- We propose Unbiased Dynamic Multimodal Learning (UDML), an architecture-agnostic framework explicitly addressing both quality estimation bias and dual-suppression bias in dynamic fusion, collectively ensuring robust dynamic multimodal learning.
- Extensive experiments on diverse multimodal benchmarks, demonstrating that UDML consistently improves performance across various tasks and settings.

2. Related Work

2.1. Dynamic Multimodal Learning

Researchers have proposed a series of dynamic multimodal fusion algorithms, which can be categorized into two classes: prior-based methods [9, 20] and uncertainty-based methods [8, 11, 16, 47].

Prior-based fusion methods allocate modality weights based on human knowledge. For example, the RGB modality typically contains more information than the infrared (IR) modality under normal lighting conditions. However, this relationship can reverse in low-light scenarios, where infrared imaging becomes more reliable. To address this, Guan et al. [9] introduce an illumination-aware fusion module that dynamically adjusts modality contributions based on the scene's lighting intensity. In addition to environmental factors, intrinsic properties of network features can also inform fusion decisions. For instance, Li et al. [20] leverage the scaling factors in batch normalization as a feature selection metric to adjust the contribution of different modalities.

Uncertainty-based fusion methods adjust modality contributions based on prediction uncertainty, offering a more theoretically grounded approach compared to prior-based methods. A commonly used uncertainty metric is predictive entropy computed from the classifier logits [11, 35, 36, 38]. However, since incorrect predictions may still yield high-confidence scores, relying solely on classifier outputs may lead to overconfidence. Thus, feature-space modeling with multivariate Gaussian distributions has been proposed to more accurately capture modality uncertainty by analyzing feature variance. This approach has seen extensive application in emotion recognition [16, 31, 37] and image-text classification [8, 47]

Despite their success, most existing uncertainty estimation methods rely on empirical heuristics, failing to assess the modality quality when noise levels are either too low or too high. More importantly, current studies assume that the initial contribution of each modality is equal. This overlooks the modality dependency imbalances induced by optimization bias. Consequently, the performance of dynamic fusion

could be inferior to that of stable fusion.

2.2. Imbalanced Multimodal Learning

Recent studies pointed out that most multimodal learning methods fail to enhance performance significantly, even with more information [6, 7, 27, 32, 43, 48]. Wang *et al*. [32] observed that different modalities exhibit varying convergence rates, leading to multimodal models that fail to surpass their unimodal counterparts. Peng *et al*. [27] further showed that the modality with superior performance tends to dominate the optimization process, leading to inadequate feature learning in weaker modalities. To this end, various methods have been developed to enhance the conventional multimodal learning framework and can be roughly categorized into two types: gradient modulation and alternating optimization. Gradient modulation [7, 27, 43] aims to enlarge the gradient of weaker modality in multimodal learning, balancing the optimization of different modality encoders. Alternating optimization methods [15, 40, 42, 48] transform the conventional joint multimodal learning process into an alternating unimodal learning process to minimize inter-modality interference directly.

However, these methods address only the underoptimization problem of modality encoders during the training stage and do not correct the model's dependence bias on each modality during inference.

3. Methods

3.1. Re-analyze the Dynamic Multimodal Learning

Notation. In a general setting, we consider two input modalities denoted as m_1 and m_2. The dataset is represented as $\mathcal{D} = \{x_i^{m_1}, x_i^{m_2}, y_i\}_{i=1,2,...,N}$, where $y \in \{1, 2, \ldots, K\}$ denotes the class labels and K is the total number of classes. We then employ two encoders $\varphi_1(\boldsymbol{\theta}_1, \cdot)$ and $\varphi_2(\boldsymbol{\theta}_2, \cdot)$ to extract features, where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ represent the parameters of each respective encoder. The feature representations are given by $\boldsymbol{z}_i^{m_1} = \varphi_1(\boldsymbol{\theta}_1, x_i^{m_1})$ and $\boldsymbol{z}_i^{m_2} = \varphi_2(\boldsymbol{\theta}_2, x_i^{m_2})$. The representations from the two encoders are typically fused via a specific fusion method, which is a common practice in multimodal learning [5, 13]. We denote the fusion module as $\phi_\tau(\boldsymbol{\theta}_\tau, \cdot)$, where $\boldsymbol{\theta}_\tau$ represents the parameters of this module. The final classification is performed by a linear classifier parameterized by $\boldsymbol{W} \in \mathbb{R}^{M\times(d_1+d_2)}$ and $\boldsymbol{b} \in \mathbb{R}^M$, and the model output for an input x_i can be expressed as follows:

$$\begin{cases} f(x_i) = \boldsymbol{W}\boldsymbol{z_i}^\tau + \boldsymbol{b} & (1) \\ \boldsymbol{z_i}^\tau = \phi_\tau(\boldsymbol{\theta_\tau}, \boldsymbol{z}_i^{m_1}; \boldsymbol{z}_i^{m_2}). & (2) \end{cases}$$

Dynamic Multimodal Learning (DML). DML considers the problem that the capacity of $z_i^{m_1}$ and $z_i^{m_2}$ varies with the input samples. Consequently, the multimodal model should adjust the decision dependency on each modality dynamically according to their representation capacity. Thus, the fusion representation of DML can be written as follows,

$$z_i{}^{\tau-dml} = \varphi_\tau(\theta_\tau, M{*}w_i{}^{m_1}{*}z_i{}^{m_1}, M{*}w_i{}^{m_2}{*}z_i{}^{m_2}), \quad (3)$$

where $w_i{}^{m_1} + w_i{}^{m_2} = 1$; M is the number of modalities, which is used to compensate for the amplitude suppression from the weight. Here, $\frac{w_i{}^{m_1}}{w_i{}^{m_2}}$ is proportional to the representation capacity of modality m_1 and m_2, assigning the low-capacity modality a lower attention.

Therefore, the accurate measure of modality capacity is the key of DML. Most existing methods quantify the modality capacity via statistical uncertainty measures $s(.)$,such as energy fraction [47] and probabilistic embedding [29],

$$w_i{}^{m_1} = g(\frac{1}{s(z_i{}^{m_1})}), \quad (4)$$

where $g(.)$ is the normalization operator. Lower uncertainty means higher capacity [8, 47].

Limitation Analysis. Existing methods inherently assume a stable and monotonic relationship between $s(z)$ and modality uncertainty. In practice, this assumption does not hold. Under low noise, $s(z)$ is insensitive to subtle quality differences, leading to nearly fixed weights. Under high noise, representation collapse distorts $s(z)$, producing inflated or misleading uncertainty estimates. Taking the widely used probabilistic embedding (PE) as an example (Fig. 1(a)), PE fails to detect low-intensity noise ($\sigma < 4$), resulting in rigid modality weighting. More critically, even under extreme corruption ($\sigma > 10$), PE still assigns considerable weight to the corrupted modality rather than disregarding it.

Besides, existing methods generally assume that the initial contributions of all modalities are identical, overlooking the intrinsic dependency bias. Specifically, we denote the intrinsic dependency of modality m as α^m, satisfying $\sum_m \alpha^m = M$. We can rewrite $z_i{}^{\tau-dml}$ as follows:

$$z_i{}^{\tau-dml} = \varphi_\tau(\theta_\tau, M{*}w_i{}^{m_1}{*}\alpha^{m_1}{*}z_i{}^{m_1}, M{*}w_i{}^{m_2}{*}\alpha^{m_2}z_i{}^{m_2}). \quad (5)$$

Ideally, $\alpha^{m_1} = \alpha^{m_2} = 1$ in a balanced multimodal system. However, multimodal models tend to rely more on easy-to-learn modalities to make decisions [17, 27], leading to an unbalanced dependency. Taking the audio-visual task as an example, assuming that audio modality is m_1 and visual modality is m_2, we can obtain the following inequality:

$$\alpha^{m_1} > 1 > \alpha^{m_2}. \quad (6)$$

This is because audio modality is easier to learn than the visual one [17, 27]. This will cause the dual suppression problem when the visual modality with high uncertainty. The hard-to-learn visual modality is first suppressed by optimization bias (low α^{m_2}), and then further down-weighted by uncertainty-based reweighting (low $w_i{}^{m_2}$). Consequently, as illustrated in Fig. 1(b), the performance of dynamic fusion could be inferior to that of static fusion.

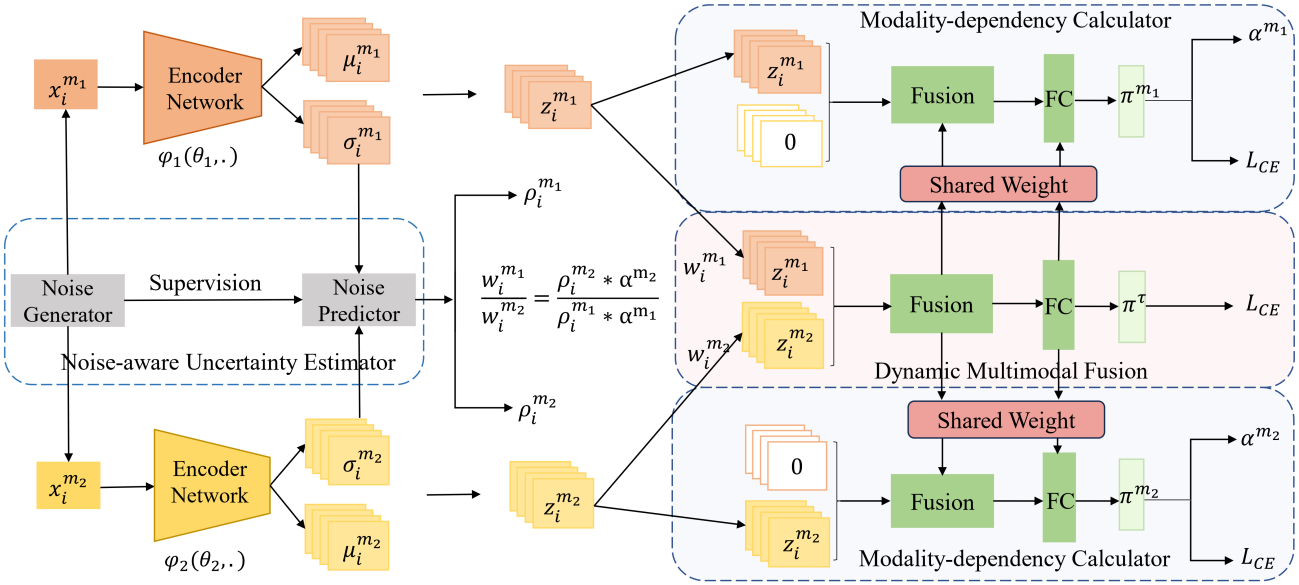

Figure 2. The framework of the unbiased dynamic multimodal fusion. It consists of two parts: 1) noise-aware uncertainty estimator, which measures the modality quality; 2) modality-dependency calculator, which quantifies the model's dependency on each modality.

3.2. Unbiased Dynamic Multimodal Learning

Accordingly, conventional dynamic multimodal learning suffers from the biases of uncertainty estimation and dual suppression. To overcome these challenges, we propose an unbiased dynamic multimodal learning framework. As shown in Fig 2, it consists of two key components: a noise-aware uncertainty estimator to quantify modality uncertainty (i.e., ${\rho_i}^{m_1}$ and ${\rho_i}^{m_2}$) as well as a modality-dependency calculator to quantify modality effect on outputs(i.e., ${\alpha}^{m_1}$ and ${\alpha}^{m_2}$). Then we can get the unbiased weight ${w_i}^{m_1}$,${w_i}^{m_2}$ as follows:

$$\frac{{w_i}^{m_1}}{{w_i}^{m_2}} = \frac{\frac{1}{{\rho_i}^{m_1} * \alpha^{m_1}}}{\frac{1}{{\rho_i}^{m_2} * \alpha^{m_2}}}. \tag{7}$$

Beyond conventional dynamic multimodal learning, the unbiased weight considers the modality-dependency bias of each modality, avoiding the problem of dual suppression. Finally, since the joint training of multimodal representations, uncertainty estimation, and the dependency calculator may be unstable, we design a progressive optimization strategy to ensure convergence.

3.2.1. Noise-aware Uncertainty Estimator

Existing uncertainty-based methods often fail to assess the modality quality when noise levels are either too low or too high. To this end, we introduce a noise-aware uncertainty estimator that learns to predict the intrinsic noise intensity of each modality through controlled perturbations.

Take the modality m_1 as an example, let $p(\sigma)$ denote a discrete set of noise levels used for perturbation. For each sample ${x_i}^{m_1}$ and noise level $\sigma \sim p(\sigma)$ we draw perturbations $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ and supervise the estimator $E(\cdot)$ to predict the scalar intensity σ. The training objective is

$$\mathcal{L}_{\text{est}} = \left\| E({x_i}^{m_1} + \epsilon_{i,k}) - \sigma_k \right\|_2^2, \tag{8}$$

where $\sigma_k \sim p(\sigma)$, and $\epsilon_{i,k} \sim \mathcal{N}(0, \sigma_k^2 I)$.

This formulation forces the model to learn a clear correspondence between feature corruption and noise level, allowing accurate uncertainty measure across both low- and high-noise conditions.

To avoid overfitting to the injected perturbation and improve robustness to unseen noise types, we do not estimate noise intensity directly from the raw perturbed input $x + \epsilon$. Instead, we adopt a probabilistic representation that separates semantic content and noise characteristics. Specifically, each modality input is encoded into a Gaussian-distributed representation:

$$z \sim \mathcal{N}(\mu(\varphi_1(\boldsymbol{\theta}_1, {x_i}^{m_1})), \Sigma(\varphi_1(\boldsymbol{\theta}_1, {x_i}^{m_1})), \tag{9}$$

where $\mu(.)$ captures semantic information and $\Sigma(.)$ models uncertainty. Here, the noise estimator receives the variance term rather than the raw signal:

$$\mathcal{L}_{\text{est}} = \left\| E(\Sigma(\varphi_1(\boldsymbol{\theta}_1, {x_i}^{m_1} + \epsilon)) - \sigma_k \right\|_2^2. \tag{10}$$

This design ensures that the estimator learns to infer noise intensity from intrinsic distributional uncertainty instead of relying on low-level pixel distortions, thereby improving generalization to diverse noise patterns.

Then the inference uncertainty is defined as follows:

$$\rho_i{}^{m_1} = E^2(\Sigma(\varphi_1(\boldsymbol{\theta}_1, x_i{}^{m_1})). \tag{11}$$

In practice, the estimator $E(\cdot)$ is implemented as a lightweight two-layer MLP operating, which introduces negligible computational overhead while providing strong robustness and noise-level calibration capability.

3.2.2. Modality-dependency Calculator

To quantify each modality's contribution to the final prediction, the modality-dependency calculator measures the sensitivity of the model's prediction to the removal of each modality. The core idea is that if dropping a modality causes a larger change in the prediction, then the model relies more on this modality.

Specifically, let π^{τ} denotes the fused logit output from all modalities, and π^{m_2} (π^{m_1}) is the logit obtained when modality m_1 (m_2) is dropped. The dependency scores d^{m_1} and d^{m_2} that quantify the contribution of each modality to the final prediction are computed as:

$$d^{m_1} = \|\pi^{\tau} - \pi^{m_2}\|_1, \quad d^{m_2} = \|\pi^{\tau} - \pi^{m_1}\|_1. \tag{12}$$

These scores are then normalized to produce scalar modality weights:

$$\alpha^{m_1} = M * \frac{d^{m_1}}{d^{m_1} + d^{m_2}}, \quad \alpha^{m_2} = M * \frac{d^{m_2}}{d^{m_1} + d^{m_2}}, \tag{13}$$

where M=2, meaning the number of modalities. α^{m_1} and α^{m_2} reflect the relative importance of each modality and can be used to reweight modality features in the fusion module adaptively. This mechanism allows the model to mitigate dual-suppression effects caused by initial dependency bias and uncertainty-based weighting, ensuring a more balanced and robust multimodal integration.

Notably, modality dropout variants reuse the same fusion and classification modules as the full multimodal branch. No additional parameters or network copies are created. Therefore, the modality-dependency computation introduces negligible overhead.

3.2.3. Progressive Optimization Strategy

To effectively train the UDML framework without inducing gradient conflicts between multiple objectives, we introduce a progressive optimization strategy. The core idea is to gradually incorporate different learning targets while preserving the stability of the primary task. Concretely, the training proceeds in two stages: multimodal representation pre-training and noise-aware training.

Stage 1: Multimodal representation pre-training. The model is first trained on clean, unperturbed data to optimize the main task (e.g., classification) and learn stable multimodal representations. The loss in this stage is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{uni}}, \tag{14}$$

where $\mathcal{L}_{\text{task}} = \mathcal{L}_{CE}(f(x_i), y)$ denotes the multimodal loss for the task at hands; $\mathcal{L}_{\text{uni}} = \mathcal{L}_{CE}(f(x_i{}^{m_1}), y) + \mathcal{L}_{CE}(f(x_i{}^{m_2}), y)$ denote the unimodal loss for modality-dependency calculator.

Stage 2: Noise-aware training. After the primary encoder is stabilized, controlled perturbations are introduced to the inputs for training the noise-aware uncertainty estimator. The loss in this stage is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{uni}} + \mathcal{L}_{\text{est}}. \tag{15}$$

Importantly, the gradient from the noise estimation loss is blocked from backpropagating into the modality encoders, preventing conflicts with the main task optimization and avoiding performance degradation. This progressive scheme allows the UDML framework to jointly learn multimodal representations, noise estimation, and the primary task within a standard training schedule while ensuring stable convergence and robust task performance.

4. Experiments

4.1. Experimental Settings

Datasets. We conduct extensive evaluations across five multimodal benchmarks with diverse modality combinations. Specifically, CMU-MOSI [45] and CMU-MOSEI [46] are tri-modal sentiment analysis datasets containing text, visual, and audio modalities. MVSA-Single [25] is a bimodal image–text dataset for multimodal sentiment analysis. CREMA-D [2] and Kinetics-Sounds (KS) [1] provide bimodal audio–visual data for emotion recognition and action understanding, respectively. This diverse set of datasets allows us to comprehensively validate our method under heterogeneous modality configurations.

Implementation Details. For the tri-modal video datasets CMU-MOSI and CMU-MOSEI, we adopt the same feature extractors as prior dynamic multimodal learning works [8, 47]: FACET for visual, COVAREP for acoustic, and BERT for textual modalities. For the image-text dataset MVSA-Single, we also adopt the same feature extractors as prior works [8, 47], ResNet-152 for RGB images, and BERT for text inputs. All models are optimized using Adam with an initial learning rate of 1×10^{-5}, a mini-batch size of 16, and we apply a Reduce-on-Plateau scheduler for learning rate adjustment.

For the CREMA-D and Kinetics-Sounds datasets, we employ ResNet18 as the backbone encoder, consistent with prior works [7, 27]. For the CREMA-D dataset, a single frame is selected from each video clip and resized to

Table 1. Performance comparison of different methods on multimodal classification tasks on MVSA-Single, CREMA-D, and Kinetics-Sounds datasets.

Methods	MVSA-Single		CREMA-D		Kinetics-Sounds	
	Acc	F1	Acc	F1	Acc	F1
Late Fusion[26]	76.88	75.72	58.83	59.43	64.97	65.21
MMBT[49]	78.50	-	64.25	65.43	65.89	-
TMC[11]	76.06	74.55	65.43	66.10	66.12	66.51
ITIN[33]	75.19	74.97	-	-	-	-
MVCN[41]	76.06	74.55	-	-	-	-
QMF[47]	78.07	77.18	66.92	67.02	66.81	67.25
EAU[8]	79.15	78.36	67.61	67.90	68.05	68.36
LDDU [16]	79.71	79.12	68.56	68.77	68.95	69.27
UDML	**80.79**	**80.10**	**70.02**	**70.58**	**69.98**	**70.26**

Table 2. Performance comparison of different methods for multimodal regression tasks on CMU-MOSI and CMU-MOSEI datasets.

Methods	CMU-MOSI			CMU-MOSEI		
	Acc7	F1	Corr	Acc7	F1	Corr
MIB[51]	48.6	85.3	0.798	54.1	86.2	0.790
HMA[23]	45.3	85.6	0.782	52.8	85.4	0.787
MIM[10]	47.0	85.9	0.805	52.5	86.3	0.792
GCNet[4]	44.9	85.1	-	51.5	85.2	-
ConFEDE[44]	42.3	85.5	0.784	54.9	85.8	0.780
DiCMoR[34]	45.3	85.6	-	53.4	85.1	-
DMD[21]	45.6	86.0	-	54.5	86.6	-
EAU[8]	48.8	86.2	0.809	54.8	86.9	0.816
LDDU [16]	49.2	86.4	0.814	55.2	87.1	0.821
UDML	**50.3**	**86.8**	**0.825**	**56.3**	**88.0**	**0.832**

224×244 to serve as the visual input, while the corresponding audio is converted into a spectrogram of size 257×299 using librosa [24]. In the case of the Kinetics-Sounds dataset, three frames are uniformly sampled from each video clip and resized to 224×224 for visual input. The entire audio data is transformed into a spectrogram with dimensions of 257×1,004. Training is carried out with configurations consistent with previous works [7, 27], including a mini-batch size of 64, an SGD optimizer with momentum set to 0.9, a learning rate of 1e-3, and a weight decay of 1e-4.

Evaluation Metric. For the classification tasks on MVSA-Single, CREMA-D, and Kinetics-Sounds, we report accuracy (Acc) and F1 score. For the regression-style sentiment analysis tasks on CMU-MOSI and CMU-MOSEI, we follow prior work [8, 47] and report 7-class accuracy (Acc7), F1 score, and Pearson correlation coefficient (Corr).

Comparison settings. We compared UDML with existing dynamical multimodal learning methods on the five datasets, including MIB [51], HMA [23], MIM [10], GCNet [4], ConFEDE [44], DiC-MoR [34], DMD [21], and QMF [47]. Moreover, we also make fair comparisons with the recent SOTA methods, including EAU [8] and LDDU [16]. Particularly, similar to the counterpart method EAU [8], we evaluate our method on noisy datasets to observe the model robustness.

Notably, despite UDML needing a two-stage optimization process, the total number of training epochs is consistent with that of existing methods, ensuring a fair comparison. Specifically, Stage 1 lasts for half of the total epochs, while Stage 2 lasts for the remaining half.

4.2. Experimental Results

Performance and comparison on multimodal classification tasks. Table 1 reports the multimodal recognition results of our proposed UDML method compared to a range of baselines on the MVSA-Single, CREMA-D, and Kinetics-Sounds datasets. As shown, UDML consistently achieves the best performance across all datasets, outperforming both static fusion methods and dynamic fusion approaches. Particularly, compared with the recent state-of-the-art method LDDU, UDML achieves at least 1% absolute improvements on all datasets. These results highlight UDML's effectiveness in dynamic multimodal fusion.

Performance and comparison on multimodal regression tasks. Table 2 presents the results of UDML and its competitors on the CMU-MOSI and CMU-MOSEI datasets

Table 3. Performance comparison of different methods when 50% of the modalities suffer from salt and Gaussian noise, respectively. The mean and variance of the noise are 0 and ϵ, respectively.

Datasets	Methods	Clean	Salt		Gaussian	
		Acc@ϵ=0	Acc@ϵ=5	Acc@ϵ=10	Acc@ϵ=5	Acc@ϵ=10
MVSA-Single	Late fusion [26]	76.88	67.88	55.43	63.46	55.16
	MMBT[49]	78.50	74.07	51.26	71.99	55.35
	TMC[11]	74.88	68.02	56.62	66.72	60.36
	QMF [47]	78.07	73.90	60.41	73.85	61.28
	EAU [8]	79.15	74.81	61.04	73.89	62.04
	LDDU [16]	79.71	75.36	61.36	74.10	62.25
	UDML	**80.79**	**78.26**	**64.52**	**77.87**	**63.62**
CREMA-D	Late fusion[26]	58.83	50.99	48.75	49.49	47.00
	MMBT[49]	64.25	55.27	52.98	54.32	51.75
	TMC[11]	65.43	58.86	54.22	56.93	53.37
	QMF[47]	66.92	59.56	55.45	58.03	54.21
	EAU[8]	67.61	60.02	56.33	59.59	55.57
	LDDU[16]	68.56	61.25	57.12	60.32	56.14
	UDML	**70.02**	**68.32**	**64.48**	**67.53**	**62.96**

for multimodal regression tasks. UDML also achieves the best performance across all metrics and both datasets. Specifically, On CMU-MOSI, UDML obtains 50.3% Acc7, 86.8% F1, and 0.825 Corr, outperforming prior representative methods such as EAU and LDDU. Similarly, on CMU-MOSEI, UDML achieves 56.3% Acc7, 88.0% F1, and 0.832 Corr, marking a consistent improvement over existing approaches. Notably, UDML maintains strong performance on both small-scale (MOSI) and large-scale (MOSEI) datasets, demonstrating robustness and generalization in multimodal sentiment regression.

Performance and comparison on noisy multimodal datasets. To assess the effectiveness of UDML in handling data noise, we follow QMF [47] and conduct more evaluation when 50% of the multimodal samples are corrupted by salt or Gaussian noise with different intensities. As shown in Table 3, performance degrades across all methods as noise intensity increases. However, UDML consistently achieves the best performance across all noise settings on both MVSA-Single and CREMA-D. Compared with existing dynamic fusion methods (e.g., EAU, LDDU), UDML shows significantly smaller performance degradation as noise intensity increases, demonstrating strong robustness to modality corruption. This improvement stems from the noise-aware uncertainty estimator providing calibrated and noise-agnostic uncertainty estimation, which enables reliable down-weighting of heavily corrupted modalities.

Importantly, although the noise estimator is trained using only Gaussian perturbations, UDML generalizes robustly to salt noise as well, confirming that the estimator does not overfit to noise type and instead learns a noise-agnostic measure of corruption. This cross-noise generalization ability indicates that the proposed variance-based uncertainty modeling successfully captures universal corruption cues shared across different noise distributions.

Table 4. Ablation of the UDML method when 50% of the modalities suffer from Gaussian noise ($\epsilon = 5$). We report the result of complete UDML (Full), the UDML without noise-aware uncertainty estimator (-NUE), the UDML without Modality-dependency Calculator (-MC), and the UDML without Progressive Optimization Strategy(POS).

Dataset	MVSA	CREMA-D	Kinetics-Sounds
Settings	Acc	Acc	Acc
Full	**77.87**	**67.53**	**66.36**
- NUE	74.56	64.60	64.33
- MC	75.44	64.10	65.41
- POS	76.16	66.49	64.92

Moreover, the improvement is particularly significant on CREMA-D. This is because CREMA-D exhibits a large modality bias where the audio modality dominates the optimization and has a higher contribution to the logit output. Conventional dynamic fusion strategies further amplify this bias when noise is present, leading to dual suppression of the visual modality. In contrast, our dependency realignment mechanism explicitly counteracts such imbalance by correcting the modality contribution shift, ensuring that both modalities remain effectively utilized. As a result, UDML achieves more stable and superior multimodal collaboration, especially in scenarios with strong modality asymmetry

4.3. Ablation Study

The ablation study of UDML. We conduct experiments to study the impact of the noise-aware uncertainty estimator (NUE), the Modality-dependency Calculator (MC), and

Table 5. Performance evaluation after integrating the UDML method into CNN-based and Transformer-based architectures. † indicates that the UDML is applied.

Datasets	MVSA	CREMA-D	Kinetics-Sounds
Structure	Acc	Acc	Acc
MMTM	78.86	51.86	56.70
MMTM†	**85.68**	**56.88**	**62.72**
mmFormer	79.69	59.69	64.72
mmFormer†	**85.33**	**63.14**	**67.51**

the Progressive Optimization Strategy (POS) on the performance of UDML. As shown in Table 4, removing any component leads to performance degradation, confirming the necessity of each part. Excluding the noise-aware uncertainty estimator (–NUE) notably reduces accuracy (e.g., 77.87→74.56 on MVSA-Single), indicating that heuristic uncertainty estimation fails to correctly assess noise levels. Removing the Modality-dependency Calculator (–MC) causes the most drop, verifying that addressing modality bias is crucial to avoid dual suppression of difficult modalities. Meanwhile, removing the Progressive Optimization Strategy (–POS) also hurts performance due to gradient conflict during joint optimization. Overall, the full UDML achieves the best performance, demonstrating that accurate noise estimation, dependency realignment, and stable optimization jointly contribute to unbiased dynamic multimodal learning.

Generalization to different multimodal architectures. To demonstrate the generality of UDML in various settings, we integrate it with two representative intermediate fusion methods, MMTM [19] and mmFormer [50], and evaluate the performance across multiple datasets. MMTM is a CNN-based architecture that fuses multimodal intermediate features using squeeze-and-excitation operations, while mmFormer is a Transformer-based model employing cross-attention for feature fusion. Specifically, UDML is inserted before the fusion block to regulate the reweighted multimodal representation. The training schedules and data augmentation strategies are kept identical to the respective baseline implementations to guarantee a fair comparison.

As shown in the right part of Table 5, although MMTM and mmFormer already achieve competitive performance without UDML, the introduction of UDML also brings significant performance gains. Specifically, on the MVSA-Single dataset, the performance of MMTM improves from 78.86% to 85.68% when the UDML is applied. These results confirm the robustness and versatility of UDML across different architectures.

Visualization. To understand the mechanism of UDML intuitively, we visualize the inference coefficients of each modality on the CREMA-D dataset when different noises are added to the image modality. As shown in Fig 3(a), the weight for visual modality (red curve) smoothly decreases

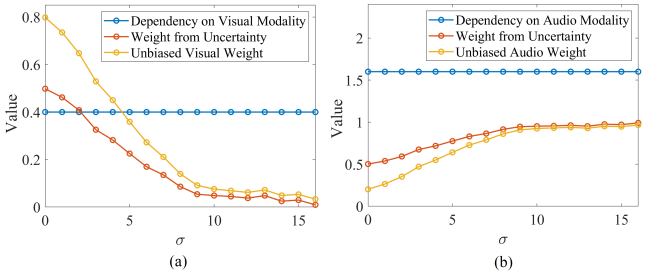

Figure 3. Visualization of dynamic multimodal fusion for audio-visual classification on the CREMA-D dataset as varying levels of noise (σ) are injected into the visual modality.

as the visual noise increases, and eventually approaches zero when the noise is large. This demonstrates that the proposed noise-aware uncertainty estimator can accurately distinguish both low-noise and high-noise cases, overcoming the misestimation problem of traditional uncertainty estimation methods.

Meanwhile, we can see that the model assigns a stronger intrinsic dependency to the audio modality (=1.6) than to the visual modality (=0.4), regardless of noise levels. This demonstrates the dependency bias of the multimodal model. By integrating both the noise-aware weighting and the modality dependency, the final unbiased fusion coefficient corrects for this reliance bias. Specifically, UDML amplifies the visual weight by 1.6× and suppresses the audio weight by 0.4×. For example, the final weights become approximately 0.8:0.2 (visual: audio) instead of 0.5:0.5 when the noise is 0. This ensures that fusion respects the intrinsic discriminative contributions while remaining robust to noise, achieving a balanced and reliable multimodal integration.

5. Conclusion

In this paper, we identify two fundamental challenges in multimodal fusion: unreliable modality-uncertainty estimation under extreme noise and the dual suppression effect, in which low-contribution modalities are further down-weighted by high uncertainty. As a result, dynamic fusion may even underperform static fusion. To address this, we propose UDML, a dynamic multimodal learning framework that combines reliable uncertainty estimation with dependency-aware fusion. Visual analysis shows that our uncertainty estimator accurately tracks modality degradation and enables adaptive suppression of corrupted modalities. Meanwhile, the derived reliance coefficients expose intrinsic contribution imbalance, which is effectively corrected by our unbiased weighting strategy. Extensive experiments on multiple multimodal datasets demonstrate that UDML yields stronger robustness, more stable fusion behavior, and better overall performance than existing fusion methods.

6. Limitation

UDML mainly focuses on modality-level bias and does not explicitly address sample-level bias. For example, tail-class samples may present high uncertainty because of limited training data rather than modality noise. In such cases, the model may over-suppress these samples by assigning unnecessarily low modality weights. This suggests that uncertainty in multimodal learning may arise not only from modality corruption but also from data imbalance and sample difficulty. In future work, we plan to investigate how to disentangle these sources of uncertainty and integrate sample-level bias correction into dynamic multimodal fusion.

Acknowledgments

This work is supported by Ministry of Science and Technology of Sichuan Province Program (2026NSFSC0430).

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 5

[2] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 5

[3] Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7088–7097, 2021. 1

[4] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Global context networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6881–6895, 2020. 6

[5] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, Long Beach, CA, USA, 2019. 3

[6] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021. 3

[7] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023. 3, 5, 6

[8] Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26876–26885, Seattle, 2024. 1, 2, 3, 5, 6, 7

[9] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019. 1, 2

[10] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Punta Cana, 2021. 6

[11] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *Proceedings of the International Conference on Learning Representations*, pages 1–10, Addis Ababa, 2020. 1, 2, 6, 7

[12] Danfeng Hong, Jingliang Hu, Jing Yao, Jocelyn Chanussot, and Xiao Xiang Zhu. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:68–80, 2021. 1

[13] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. Deep multimodal multilinear fusion with high-order polynomial pooling. Vancouver, Canada, 2019. 3

[14] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019. 1

[15] Cong Hua, Qianqian Xu, Shilong Bao, Zhiyong Yang, and Qingming Huang. Reconboost: Boosting can achieve modality reconcilement. *arXiv preprint arXiv:2405.09321*, 2024. 3

[16] Jingwang Huang, Jiang Zhong, Qin Lei, Jingpeng Gao, Yuming Yang, Sirui Wang, Peiguang Li, and Kaiwen Wei. Latent distribution decoupling: A probabilistic framework for uncertainty-aware multimodal emotion recognition. In *Proceedings of the 2025 ACL Findings*, Vienna, Austria, 2025. 1, 2, 6, 7

[17] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International Conference on Machine Learning*, pages 9226–9259. PMLR, 2022. 2, 3

[18] Wen-Da Jin, Jun Xu, Qi Han, Yi Zhang, and Ming-Ming Cheng. Cdnet: Complementary depth network for rgb-d salient object detection. *IEEE Transactions on Image Processing*, 30:3376–3390, 2021. 1

[19] 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*, pages 13289–13299, Seattle, Washington, 2020. 8

[20] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018. 1, 2

[21] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640, Vancouver, Canada, 2023. 6

[22] Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, Zichang Tan, Qi Yuan, Kai Wang, Chi Lin, Guodong Guo, Isabelle Guyon, et al. Multi-modal face anti-spoofing attack detection challenge at cvpr2019. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[23] Xiao Liu, Weimin Li, Shang Miao, Fangyu Liu, Ke Han, and Tsigabu T Bezabih. Hammf: hierarchical attention-based multi-task and multi-modal fusion model for computer-aided diagnosis of alzheimer's disease. *Computers in Biology and Medicine*, 176:108564, 2024. 6

[24] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015. 6

[25] Tianrui Niu, Shuai Zhu, Liang Pang, et al. Sentiment analysis on multi-view social data. In *MultiMedia Modeling: International Conference*, pages 15–27, Miami, FL, USA, 2016. 5

[26] Su Pang, Daniel Morris, and Hayder Radha. Clocs: Camera-lidar object candidates fusion for 3d object detection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 10386–10393, Las Vegas, Nevada, 2020. IEEE. 6, 7

[27] Xiaokang Peng, Yake Wei, Andong Deng, et al. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022. 2, 3, 5, 6

[28] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13525–13531. IEEE, 2021. 1

[29] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019. 3

[30] Peng Sun, Wenhu Zhang, Huanyu Wang, Songyuan Li, and Xi Li. Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1407–1417, 2021. 1

[31] M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):805–822, 2023. 2

[32] 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*, pages 12695–12705, 2020. 3

[33] Xiaoling Wang, Qi Kang, MengChu Zhou, Siya Yao, and Abdullah Abusorrah. Domain adaptation multitask optimization. *IEEE Transactions on Cybernetics*, 53(7):4567–4578, 2022. 6

[34] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 22025–22034, Paris, France, 2023. 6

[35] Shicai Wei, Chunbo Luo, and Yang Luo. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20039–20049, 2023. 2

[36] Shicai Wei, Chunbo Luo, and Yang Luo. Scaled decoupled distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15975–15983, 2024. 2

[37] Shicai Wei, Yang Luo, Yuji Wang, and Chunbo Luo. Robust multimodal learning via representation decoupling. In *European conference on computer vision*, pages 38–54. Springer, 2024. 2

[38] Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2250–2259, 2025. 2

[39] Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22879–22888, 2025. 1

[40] Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. *arXiv preprint arXiv:2405.17730*, 2024. 3

[41] Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 5240–5252, Toronto, Canada, 2023. 6

[42] Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. In *European Conference on Computer Vision*, pages 71–86. Springer, 2025. 3

[43] Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[44] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630, Toronto, Canada, 2023. 6

[45] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2018. 5

[46] 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*, pages 2236–2246, Melbourne, Australia, 2018. 5

[47] Qian Zhang, Huixuan Wu, Chuhan Zhang, et al. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, pages 41753–41769. PMLR, 2023. 1, 2, 3, 5, 6, 7

[48] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. *arXiv preprint arXiv:2311.10707*, 2023. 3

[49] Yutong Zhang, Qingyang Li, and Shuang Zhao. Multimodal transformer for medical image analysis. *arXiv preprint arXiv:2105.12345*, 2021. 6, 7

[50] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. *arXiv preprint arXiv:2206.02425*, 2022. 8

[51] Shuang Zhao, Yutong Zhang, and Qingyang Li. Multimodal information bottleneck for representation learning. *arXiv preprint arXiv:2101.12345*, 2021. 6

[52] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Rgb-d salient object detection: A survey. *Computational Visual Media*, 7(1):37–69, 2021. 1