

Dynamic Alignment and Calibration for Multimodal Learning

Jinghao Xu¹
 nolle98xu@gmail.com
 Zhenhua Guo²
 cszguo@gmail.com
 Xiaofeng Zhu³
 seanzhuxf@gmail.com
 Xiaoshuang Shi¹
 xsshi2013@gmail.com

¹ School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
² Tianyijiaotong Technology Ltd., China
³ School of Computer Science and Technology, Hainan University, Haikou, China

Abstract

Dynamic multimodal learning aims to learn robust representations by adaptively modeling information discrepancies across modalities. However, existing methods still suffer from two limitations: (i) static cross-modal alignment strategies usually impose uniform constraints on all samples while overlooking sample-wise variations, potentially leading to unreasonable over-alignment; and (ii) confidence- or uncertainty-aware fusion methods often fail to adequately account for feature magnitude and confidence differences across modalities. For modality pairs with significant feature magnitude differences or small confidence gaps, it might be unreliable to strictly align fusion weights according to confidence. To address these issues, we propose an Alignment- and Calibration-driven Multimodal Learning framework (ACML). Specifically, ACML incorporates a dynamic cross-modal triplet alignment module, which enforces strong semantic consistency for high-confidence positive pairs while encouraging diverse representation learning between high- and low-confidence positive pairs according to their confidence gaps. Additionally, ACML introduces a difference-aware attention calibration strategy that adaptively adjusts attention regularization based on feature magnitude and confidence differences across modalities, thereby mitigating biases caused by unreasonable fusion constraints. Extensive experiments on multiple multimodal benchmark datasets demonstrate that ACML consistently achieves superior performance and robustness over recent state-of-the-art methods.

1 Introduction

Multimodal learning has achieved significant success in various fields, such as medical diagnosis[11, 27, 36, 42], emotion recognition[9, 12, 19], and scene understanding[13, 30, 33, 46], by integrating complementary information from diverse modalities to enhance representation power and generalization performance. As the pivotal mechanisms for achieving these advantages, alignment and fusion constitute two core sub-problems in multimodal

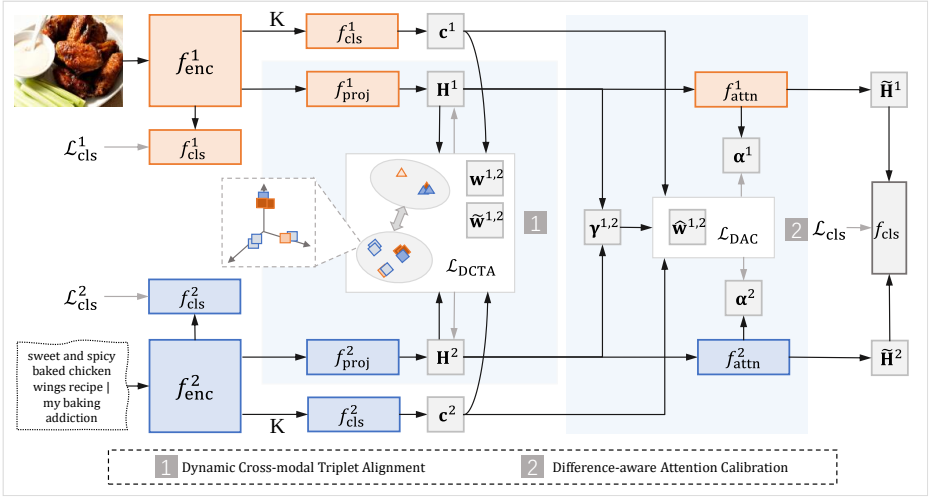


Figure 1: Overview of the proposed framework ACML. In $\mathcal{L}_{DCTA}$, $\triangle$ and $\square$ denote different categories, colors represent modalities, and color intensity reflects confidence.

learning. However, in real-world open environments, data quality and reliability across different modalities often fluctuate dynamically due to device heterogeneity, environmental noise, and occlusions[42]. This necessitates multimodal modeling that not only captures cross-modal consistency but also adapts to sample-level variations. Unfortunately, existing methods fall short of adequately modeling such dynamics in both the alignment and fusion stages.

During the alignment stage, most methods still rely on uniform static constraints, applying the same alignment intensity to all samples[43]. This paradigm overlooks the significant disparities in discriminability across different samples[49]: while strong alignment helps enhance cross-modal consistency for high-confidence samples that are easy to classify, applying the same intensity to samples with semantic ambiguity or high noise levels may be counterproductive. Forcibly compressing the distance between all positive sample pairs inherently diminishes inter-modal diversity, thereby suppressing the expression of complementary information[44, 45]. Ultimately, this leads to feature space degradation and undermines the representational advantages that multimodal models are intended to possess.

During the fusion stage, recent studies[8, 46, 47] introduce confidence- or uncertainty-aware dynamic weighting mechanisms to adaptively calibrate modality contributions. However, these methods often overlook feature magnitude differences and confidence differences across modalities. When the feature magnitude difference between modalities is large, their actual contributions to the fused representation are often already significantly differentiated. In such cases, imposing additional strong constraints may lead to unreasonable contribution reallocation and consequently interfere with effective multimodal fusion. On the other hand, when the confidence difference between modalities is small, excessively enforcing confidence-based modulation of modality contributions may restrict potential synergistic modeling across modalities.

To address the aforementioned issues, we propose an Alignment- and Calibration-driven Multimodal Learning framework (ACML), as illustrated in Figure 1. ACML consists of two

core components: a dynamic cross-modal triplet alignment (DCTA) module and a difference-aware attention calibration (DAC) module. Specifically, modality-specific features are first extracted using dedicated encoders, followed by an independent classification head with multiple stochastic forward passes for sample-level confidence estimation. The features are then projected through independent layers for dimensional alignment and jointly optimized under the proposed dynamic triplet alignment mechanism. This mechanism enhances inter-class discriminability while enforcing cross-modal semantic consistency for high-confidence positive pairs. For low-confidence samples, the alignment constraint is relaxed and an orthogonality encouragement strategy is introduced to preserve potential complementary information and mitigate noise propagation caused by forced alignment. Furthermore, in the DAC module, we impose consistency regularization on the weighted attention distributions across modalities for the same sample, while adaptively adjusting the regularization strength according to feature magnitude differences and confidence differences among modalities. This design enables more reasonable sample-aware fusion while avoiding unnecessary over-regularization on high-magnitude modalities. Extensive experimental results demonstrate that ACML achieves superior effectiveness and robustness under both standard settings and challenging scenarios involving noisy and missing modalities.

1. We propose ACML, an alignment- and calibration-driven multimodal learning framework that achieves robust and adaptive multimodal representation learning through dynamic cross-modal alignment and difference-aware attention calibration.
2. We design DCTA to enhance semantic consistency for high-confidence positive pairs while promoting diversity between high- and low-confidence pairs, thereby preserving complementary information and mitigating noise propagation.
3. We propose DAC, which dynamically adjusts the strength of attention regularization according to feature magnitude differences and confidence differences, thereby avoiding unreasonable over-regularization on attention weights.

2 Related Work

2.1 Multimodal Alignment

Multimodal alignment aims to establish semantic correspondences across different modalities (e.g., text and images) by mapping their representations into a unified shared embedding space. According to the underlying alignment mechanisms, existing approaches can be broadly categorized into explicit and implicit alignment strategies[[10](#)]. Explicit methods [[18](#), [35](#), [49](#)] typically rely on similarity matrices to directly measure cross-modal associations, whereas implicit methods[[15](#), [16](#), [20](#)] learn joint representations in a latent space through translation or prediction-based objectives. Along the methodological evolution, contrastive learning has emerged as a dominant paradigm for multimodal alignment[[18](#)]. Its core principle is to pull positive sample pairs closer while pushing negative pairs apart, thereby enforcing discriminative cross-modal consistency in the embedding space [[6](#), [49](#)]. Our method follows the explicit alignment paradigm but addresses the limited consideration of modality dynamics in prior work by introducing a dynamic-aware mechanism. Moreover, we extend conventional contrastive objectives by designing a triplet alignment loss, which aims to strike a better balance between promoting cross-modal consistency and preserving complementary information.

2.2 Multimodal Fusion

Multimodal fusion aims to integrate complementary information from different modalities to improve performance on downstream tasks. Existing approaches are typically categorized into early, intermediate, or late fusion based on the fusion stage [9]. Early fusion methods [28, 40] combine multiple modalities directly at the data level, but they often struggle with high-dimensional or heterogeneous data. Intermediate fusion approaches [6] focus on learning joint feature representations across modalities, while late fusion methods [8, 41] emphasize semantic-level integration of individual modality predictions.

Recent studies have increasingly focused on dynamic multimodal fusion [22, 37, 42, 44] to address the observation that the relative importance of each modality may vary across samples or over time. Unlike static fusion strategies that assume fixed modality weights regardless of input conditions, dynamic fusion approaches leverage attention mechanisms or uncertainty modeling to adaptively modulate the contribution of each modality. This adaptability enables them to better handle challenging scenarios, such as degraded data quality or cross-domain variations, thereby improving both robustness and generalization. However, existing studies typically impose homogeneous constraints on fusion weights while overlooking the differences in feature norms and confidence across modalities. This oversight may lead to unreasonable constraints on modality pairs with abnormal norm ratios or highly consistent confidence, thereby suppressing the expression of complementary cross-modal information. In contrast, we introduce a pairwise constraint loss that adaptively adjusts the constraint strength for different modality pairs based on both norm and confidence differences.

3 Method

3.1 Feature Extraction and Confidence Estimation

Given a mini-batch of N multimodal samples, each of which has M modality-specific inputs and a corresponding class label. The inputs are represented as $\{\mathbf{X}^m\}_{m=1}^M$, where $\mathbf{X}^m = \{\mathbf{x}_i^m\}_{i=1}^N$ denotes the m -th modality of N samples. The associated label set is $\mathbf{Y} = \{y_1, y_2, \dots, y_N\}$, with $y_i \in \{1, \dots, C\}$.

For the m -th modality, the input $\mathbf{X}^m$ is first passed through a modality-specific encoder $f_{\text{enc}}^m(\cdot)$ to extract high-level feature representations:

$$\mathbf{Z}^m = f_{\text{enc}}^m(\mathbf{X}^m), \quad (1)$$

where $\mathbf{Z}^m = [\mathbf{z}_1^m, \mathbf{z}_2^m, \dots, \mathbf{z}_N^m]^\top \in \mathbb{R}^{N \times d_m}$ is the feature matrix, $\mathbf{z}_i^m \in \mathbb{R}^{d_m}$ represents the feature vector of the i -th sample.

To obtain robust and stable confidence estimates during training, we employ a stochastic inference scheme. Specifically, we leverage the classifier $f_{\text{cls}}^m(\cdot)$ to construct two parallel computational paths. In the primary training path, the feature $\mathbf{z}_i^m$ is directly fed into the classifier without activating dropout. This path is supervised by the standard cross-entropy loss $\mathcal{L}_{\text{cls}}^m$, which is to ensure that the model learns fundamental discriminative representations:

$$\mathcal{L}_{\text{cls}}^m = -\frac{1}{N} \sum_{i=1}^N \log \left([s(f_{\text{cls}}^m(\mathbf{z}_i^m))]_{y_i} \right), \quad (2)$$

where $s(\cdot)$ denotes the softmax activation function, and $[\cdot]_{y_i}$ extracts the probability corresponding to the ground-truth index. Meanwhile, to mitigate the prevalent over-confidence issue associated with a single deterministic forward pass and to enhance the quality of confidence estimation, we introduce a probabilistic auxiliary path. Under this path, we apply a dropout operator $\delta(\cdot)$ to the same feature vector $\mathbf{z}_i^m$ and execute K stochastic forward passes. In this work, we set the dropout probability to 0.1 and the number of stochastic forward passes K to 10. For each pass $k \in \{1, \dots, K\}$, the generated stochastic prediction vector is formulated as:

$$\mathbf{p}_{i,k}^m = s(f_{\text{cls}}^m(\delta(\mathbf{z}_i^m))). \quad (3)$$

The final confidence score c_i^m for the i -th sample is defined as the empirical mean of the ground-truth probabilities across these K stochastic predictions:

$$c_i^m = \frac{1}{K} \sum_{k=1}^K [\mathbf{p}_{i,k}^m]_{y_i}. \quad (4)$$

After obtaining the confidence vector $\mathbf{c}^m = [c_1^m, c_2^m, \dots, c_N^m]^\top \in \mathbb{R}^N$, we detach it from the computational graph to serve as dynamic guidance for subsequent modules.

3.2 Dynamic Cross-modal Triplet Alignment

Cross-modal semantic alignment is a fundamental component of multimodal learning [17]. However, due to the varying discriminative capabilities across modalities and the need to preserve complementary information, enforcing uniform semantic consistency constraints on all sample pairs is often suboptimal. To address this issue, we propose a Dynamic Cross-modal Alignment loss, denoted as $\mathcal{L}_{\text{DCTA}}$. The proposed loss dynamically adjusts alignment strength according to sample confidence and explicitly preserves feature separability based on confidence discrepancy, thereby alleviating noisy semantic aggregation and promoting feature diversity.

For the feature representation $\mathbf{Z}^m$ of the m -th modality, we first employ a modality-specific projection layer $f_{\text{proj}}^m(\cdot)$, implemented as an MLP, to map the features into a shared multimodal embedding space:

$$\mathbf{H}^m = f_{\text{proj}}^m(\mathbf{Z}^m) = \{\mathbf{h}_1^m, \mathbf{h}_2^m, \dots, \mathbf{h}_N^m\}, \quad (5)$$

where $\mathbf{h}_i^m \in \mathbb{R}^d$ denotes the embedded representation of the i -th sample in the m -th modality.

Within the embedding space, we compute the cosine similarity matrix $\mathbf{S}^{m,n} \in \mathbb{R}^{N \times N}$ for each modality pair (m, n) , where each element $s_{ij}^{m,n}$ is defined as:

$$s_{ij}^{m,n} = \frac{\mathbf{h}_i^{m\top} \mathbf{h}_j^n}{\|\mathbf{h}_i^m\| \cdot \|\mathbf{h}_j^n\|}. \quad (6)$$

To dynamically regulate the alignment strength according to paired-sample reliability, we define a joint confidence weight:

$$w_{ij}^{m,n} = \min(c_i^m, c_j^n). \quad (7)$$

In this way, high-confidence positive pairs are encouraged to achieve stronger semantic consistency, while unreliable pairs are softly relaxed to alleviate noisy semantic aggregation.

The alignment term is formulated as:

$$\text{Align}_i^{m,n} = \sum_{j=1}^N \mathbb{I}(y_i = y_j) w_{ij}^{m,n} \exp\left(\frac{s_{ij}^{m,n}}{\tau}\right), \quad (8)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function and τ is a temperature parameter.

Although the confidence-weighted alignment effectively alleviates noisy aggregation caused by unreliable samples, it does not explicitly preserve the intrinsic representation diversity within the same class. To this end, we further introduce a confidence-gap-aware decorrelation constraint to encourage more diverse feature representations among positive pairs. Specifically, we define a confidence-gap weight as:

$$\tilde{w}_{ij}^{m,n} = |c_i^m - c_j^n|. \quad (9)$$

This weight adaptively modulates the strength of the decorrelation constraint according to the confidence discrepancy between positive pairs. Based on this weighting scheme, we further construct an orthogonality-inspired decorrelation objective using the absolute feature similarity. Positive pairs with larger confidence gaps are subjected to stronger decorrelation, encouraging them to maintain lower feature correlation. As a result, the learned representations remain semantically consistent while preserving sufficient diversity, thereby retaining potential complementary information. The decorrelation term is defined as:

$$\text{Decor}_i^{m,n} = \sum_{j=1}^N \mathbb{I}(y_i = y_j) \tilde{w}_{ij}^{m,n} \exp\left(\frac{|s_{ij}^{m,n}|}{\tau}\right). \quad (10)$$

For negative pairs ($y_i \neq y_j$), we further introduce a repulsion term to separate irrelevant category representations in the embedding space:

$$\text{Repel}_i^{m,n} = \sum_{j=1}^N \mathbb{I}(y_i \neq y_j) \exp\left(\frac{s_{ij}^{m,n}}{\tau}\right). \quad (11)$$

Finally, we integrate the alignment, decorrelation, and repulsion terms into a unified optimization objective. The batch-wise loss between modality m and modality n is:

$$\mathcal{L}_{\text{DCTA}}^{m,n} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\text{Align}_i^{m,n}}{\text{Align}_i^{m,n} + \text{Decor}_i^{m,n} + \text{Repel}_i^{m,n}}. \quad (12)$$

The final global objective is obtained by summing all bidirectional modality-pair losses:

$$\mathcal{L}_{\text{DCTA}} = \sum_{1 \leq m < n \leq M} (\mathcal{L}_{\text{DCTA}}^{m,n} + \mathcal{L}_{\text{DCTA}}^{n,m}). \quad (13)$$

3.3 Difference-aware Attention Calibration

In dynamic multimodal fusion, modality contributions are typically modulated by predictive confidence or uncertainty [3, 14, 15]. Existing methods typically adopt a unified global attention calibration strategy without adequately considering sample-wise modality differences. However, when the feature magnitude difference between modalities is large, their

contributions to the fused representation are often already significantly differentiated, making additional strong calibration potentially unnecessary and even harmful to effective multimodal fusion. Meanwhile, when the confidence difference between modalities is small, excessive calibration may suppress potential synergistic information across modalities. To address these issues, we propose a difference-aware attention calibration constraint, termed $\mathcal{L}_{\text{DAC}}$, which adaptively emphasizes modality pairs with distinguishable confidence levels and comparable feature magnitudes, thereby enabling more adaptive and effective multimodal fusion.

Specifically, for the i -th sample, we employ a modality-specific attention layer f_{attn}^m , implemented using a modality-specific MLP followed by a Sigmoid function, to compute the attention score for each modality. The original features are then reweighted by the corresponding attention scores to obtain the attention-weighted features $\tilde{\mathbf{h}}_i^m$, as follows:

$$\alpha_i^m, \tilde{\mathbf{h}}_i^m = f_{\text{attn}}^m(\mathbf{h}_i^m). \quad (14)$$

To calibrate the relative attention allocation at the sample level, we define the pairwise attention ratio and confidence ratio for each modality pair (m, n) as:

$$\text{AttnRatio}_i^{m,n} = \frac{\alpha_i^m}{\alpha_i^m + \alpha_i^n + \varepsilon}, \quad \text{ConfRatio}_i^{m,n} = \frac{c_i^m}{c_i^m + c_i^n + \varepsilon}, \quad (15)$$

where $\text{AttnRatio}_i^{m,n}$ measures the relative attention allocation between modalities, while $\text{ConfRatio}_i^{m,n}$ represents their relative predictive confidence. Here, ε is a small constant introduced for numerical stability.

Furthermore, to selectively enforce reliable calibration signals, we define a difference-aware pairwise weight as:

$$\hat{w}_i^{m,n} = \gamma_i^{m,n} \cdot (|c_i^m - c_i^n| + 1), \quad (16)$$

where $\gamma_i^{m,n} = \frac{\min(\|\mathbf{h}_i^m\|, \|\mathbf{h}_i^n\|)}{\max(\|\mathbf{h}_i^m\|, \|\mathbf{h}_i^n\|) + \varepsilon}$ measures the similarity between modality magnitudes. When the magnitude difference between modalities is large, the corresponding calibration constraint is adaptively relaxed to avoid unreasonable contribution reallocation. In contrast, modality pairs with comparable feature magnitudes receive relatively larger calibration weights. Meanwhile, the confidence difference term $(|c_i^m - c_i^n| + 1)$ preserves a basic alignment objective while further emphasizing modality pairs with more distinguishable reliability levels. Therefore, $\hat{w}_i^{m,n}$ jointly models feature magnitude differences and confidence differences, enabling more adaptive and reliable attention calibration.

Finally, the difference-aware attention calibration constraint between modality m and modality n is formulated as:

$$\mathcal{L}_{\text{DAC}}^{m,n} = \frac{1}{N} \sum_{i=1}^N \hat{w}_i^{m,n} (\text{AttnRatio}_i^{m,n} - \text{ConfRatio}_i^{m,n})^2. \quad (17)$$

The global difference-aware attention calibration constraint is obtained by summing over all modality pairs:

$$\mathcal{L}_{\text{DAC}} = \sum_{1 \leq m < n \leq M} \mathcal{L}_{\text{DAC}}^{m,n}. \quad (18)$$

3.4 Fusion and Optimization Objective

After cross-modal alignment and attention calibration, we adopt a mean fusion strategy to aggregate all reweighted modality representations and perform final classification through the classification head $f_{\text{cls}}(\cdot)$:

$$\mathbf{h}_i^{\text{fuse}} = \frac{1}{M} \sum_{m=1}^M \tilde{\mathbf{h}}_i^m, \quad \mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N \log \left(\left[s \left(f_{\text{cls}}(\mathbf{h}_i^{\text{fuse}}) \right) \right]_{y_i} \right). \quad (19)$$

Finally, the overall optimization objective of ACML is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{CLS}} + \lambda_1 \mathcal{L}_{\text{DCTA}} + \lambda_2 \mathcal{L}_{\text{DAC}}, \quad (20)$$

where $\mathcal{L}_{\text{CLS}} = \mathcal{L}_{\text{cls}} + \sum_{m=1}^M \mathcal{L}_{\text{cls}}^m$ denotes the summation of the fused classification loss and all unimodal classification losses, while λ_1 and λ_2 are balancing hyperparameters for $\mathcal{L}_{\text{DCTA}}$ and $\mathcal{L}_{\text{DAC}}$, respectively. For clarity, we summarize the whole training procedure of ACML in Algorithm 1.

Algorithm 1 ACML

Input: Training data $\mathcal{D}$, the number of training iterations T .

Output: Well trained ACML.

```

1: for iteration  $\in [1, T]$  do
2:   Randomly select a batch of samples  $\{\{\mathbf{X}^m\}_{m=1}^M, \mathbf{Y}\}$  from  $\mathcal{D}$ ;
3:   # Obtain modality-specific representation.
4:    $\{\mathbf{Z}^m\}_{m=1}^M \leftarrow \{f_{\text{enc}}^m(\mathbf{X}^m)\}_{m=1}^M$ 
5:   # Get the confidence score for each modal sample.
6:    $\{\mathbf{c}^m\}_{m=1}^M \leftarrow \text{Eqs. (2)-(4)}$ 
7:   # Map multimodal features to the same dimension.
8:    $\{\mathbf{H}^m\}_{m=1}^M \leftarrow \{f_{\text{proj}}^m(\mathbf{Z}^m)\}_{m=1}^M$ 
9:   # Perform dynamic cross-modal triplet alignment.
10:   $\mathcal{L}_{\text{DCTA}} \leftarrow \text{Eqs. (6)-(13)}$ 
11:  # Compute attention weights and reweight features.
12:   $\alpha^m, \tilde{\mathbf{H}}^m \leftarrow f_{\text{attn}}^m(\mathbf{H}^m)$ 
13:  # Perform difference-aware attention calibration.
14:   $\mathcal{L}_{\text{DAC}} \leftarrow \text{Eqs. (15)-(18)}$ 
15:  # Perform multimodal fusion and final classification.
16:   $\mathbf{H}^{\text{fuse}}, \mathcal{L}_{\text{cls}} \leftarrow \text{Eq. (19)}$ 
17:  # Integrate multiple constraints into the final optimization objective.
18:   $\mathcal{L} \leftarrow \text{Eq. (20)}$ 
19:  Back-propagate  $\mathcal{L}$  to update model parameters;
20: end for
```

4 Experiments and Analysis

In this section, we evaluate the effectiveness of the proposed ACML by addressing the following six Research Questions (RQs):

- **RQ1 (Effectiveness & Robustness):** How does ACML compare with SOTA baselines in terms of overall performance and robustness against noise?

- **RQ2 (Ablation Study):** What are the individual contributions of the key components within ACML to its overall performance?
- **RQ3 (Visualization Analysis):** How does $\mathcal{L}_{\text{DCTA}}$ influence the cross-modal feature distribution in the embedding space?
- **RQ4 (Hyper-parameter Analysis):** How do the critical hyper-parameters influence the stability of ACML?
- **RQ5 (Applicability to Missing Modalities):** Can ACML effectively enhance prompt-tuning frameworks as a plug-and-play module in challenging missing-modality scenarios?
- **RQ6 (Scalability to Multiple Modalities):** Does ACML maintain its effectiveness and scalability when extended to complex scenarios with more than two modalities ($M > 2$)?

4.1 RQ1 (Effectiveness & Robustness)

Setting. We evaluate ACML on two representative multimodal tasks: (1) **Scene Recognition** on the NYU Depth V2 [25] and SUN RGB-D [26] datasets, and (2) **Image–Text Classification** on the UPMC FOOD101 [32] dataset. Detailed dataset descriptions are provided in the supplementary material. All models are implemented in PyTorch and optimized using Adam on an NVIDIA A100 GPU. For the scene recognition task, we adopt ResNet-18 as the backbone network, with a learning rate of 1×10^{-4} , a batch size of 16, and an embedding dimension d of 512. For the image–text classification task, ResNet-152 and BERT are used to extract image and text features, respectively, with a learning rate of 5×10^{-5} , an effective batch size of 640 achieved via gradient accumulation (i.e., batch size 32 with 20 accumulation steps), and an embedding dimension d of 2048. Following [20], we simulate real-world multimodal uncertainty by injecting Gaussian noise into images and blank noise into text. Scene recognition experiments are repeated over 10 random seeds, while image–text classification experiments are conducted with 5 random seeds. We report both the average accuracy (with standard deviation) and the worst-case accuracy. All models are trained for up to 100 epochs, with early stopping applied to prevent overfitting.

Analysis. As shown in Table 1, we compare ACML with three categories of methods: traditional fusion approaches (**Concat** and **LateFusion**), state-of-the-art dynamic learning methods (**TMC**[27], **QMF**[41], and **PDF**[3]), and modality-imbalance learning methods (**LFM**[58] and **OPM**[54]). The results show that ACML consistently achieves the best performance across all datasets under both clean and noisy conditions. Specifically, under the clean setting ($\epsilon = 0$), ACML yields relative gains of 1.82%, 1.64%, and 0.41% in average accuracy over the strongest baseline on NYU Depth V2, SUN RGB-D, and FOOD101, respectively. As the noise level increases, the advantage of ACML becomes more pronounced on NYU Depth V2 and SUN RGB-D. For example, under severe corruption ($\epsilon = 10$), the relative gains in average accuracy further increase to 6.00% and 5.36%, respectively. This trend suggests that ACML not only improves predictive performance under clean conditions but also exhibits stronger robustness against modality corruption.

4.2 RQ2 (Ablation Study)

Setting. We conduct ablation studies on the NYU Depth V2 dataset to systematically analyze the contributions of the dynamic cross-modal triplet alignment loss ($\mathcal{L}_{\text{DCTA}}$) and the

Table 1: Classification comparison when 50% of the modalities are corrupted with Gaussian noise, i.e., zero mean with variance of ε . We highlight the optimal results using bold text and indicate the suboptimal results with underlining for clarity.

Dataset	Method	$\varepsilon = 0.0$		$\varepsilon = 5.0$		$\varepsilon = 10.0$	
		Avg $\pm$ Std	Worst	Avg $\pm$ Std	Worst	Avg $\pm$ Std	Worst
NYU Depth V2	Depth	61.97 $\pm$ 0.99	60.40	49.65 $\pm$ 3.22	43.27	42.42 $\pm$ 2.42	36.85
	RGB	62.57 $\pm$ 0.82	61.32	52.56 $\pm$ 2.28	47.55	46.16 $\pm$ 3.10	38.84
	LateFusion	69.22 $\pm$ 0.98	67.43	56.35 $\pm$ 2.84	51.22	48.65 $\pm$ 3.40	42.36
	Concat	68.70 $\pm$ 0.67	67.89	55.76 $\pm$ 2.23	49.54	48.17 $\pm$ 3.36	40.21
	TMC	<u>71.42$\pm$0.84</u>	<u>70.03</u>	61.44 $\pm$ 2.19	56.58	53.36 $\pm$ 3.36	45.87
	QMF	68.55 $\pm$ 1.29	66.82	56.60 $\pm$ 3.12	50.31	49.04 $\pm$ 3.69	41.44
	PDF	68.46 $\pm$ 0.69	66.67	56.67 $\pm$ 2.69	48.47	48.61 $\pm$ 3.56	38.23
	LFM	69.27 $\pm$ 0.62	68.20	59.17 $\pm$ 2.15	53.98	51.35 $\pm$ 2.70	44.95
	OPM	71.35 $\pm$ 0.93	69.88	63.35 $\pm$ 1.41	<u>60.40</u>	54.82 $\pm$ 2.37	<u>49.54</u>
	ACML	72.72$\pm$0.81	71.56	65.43$\pm$1.53	60.86	58.11$\pm$2.13	49.69
SUN RGB-D	Depth	52.58 $\pm$ 0.67	51.43	39.05 $\pm$ 1.06	37.15	34.66 $\pm$ 1.61	31.08
	RGB	57.46 $\pm$ 0.39	56.84	48.05 $\pm$ 0.88	45.25	42.45 $\pm$ 1.47	38.74
	LateFusion	59.90 $\pm$ 0.45	59.24	47.78 $\pm$ 2.03	43.01	41.27 $\pm$ 2.82	35.57
	Concat	59.52 $\pm$ 0.50	58.58	48.54 $\pm$ 1.63	45.12	42.74 $\pm$ 1.98	38.96
	TMC	61.02 $\pm$ 0.37	60.34	50.64 $\pm$ 1.69	47.18	43.89 $\pm$ 3.22	37.54
	QMF	61.11 $\pm$ 0.26	60.59	51.44 $\pm$ 1.06	49.13	46.45 $\pm$ 1.61	<u>42.71</u>
	PDF	60.89 $\pm$ 0.39	60.23	50.24 $\pm$ 1.43	46.71	44.92 $\pm$ 2.21	40.44
	LFM	61.98 $\pm$ 0.25	61.64	<u>52.84$\pm$0.96</u>	<u>50.55</u>	<u>47.36$\pm$1.72</u>	42.58
	OPM	<u>62.23$\pm$0.26</u>	<u>61.79</u>	51.72 $\pm$ 1.08	49.60	43.62 $\pm$ 1.90	40.59
	ACML	63.25$\pm$0.25	62.76	56.50$\pm$0.87	54.15	49.90$\pm$1.15	47.48
FOOD 101	Img	64.78 $\pm$ 0.17	64.54	34.55 $\pm$ 0.58	33.66	32.99 $\pm$ 0.28	32.54
	Text	86.49 $\pm$ 0.05	86.44	67.35 $\pm$ 0.18	67.04	43.84 $\pm$ 0.31	43.39
	LateFusion	90.71 $\pm$ 0.09	90.61	70.07 $\pm$ 2.86	64.56	59.51 $\pm$ 0.87	57.81
	Concat	88.35 $\pm$ 0.57	87.68	63.28 $\pm$ 2.35	59.45	51.18 $\pm$ 3.71	46.34
	TMC	90.26 $\pm$ 0.17	89.94	73.69 $\pm$ 0.82	72.15	61.51 $\pm$ 0.21	61.01
	QMF	92.88 $\pm$ 0.11	92.70	<u>76.39$\pm$0.62</u>	75.30	61.86 $\pm$ 0.41	61.74
	PDF	<u>92.98$\pm$0.17</u>	92.71	75.34 $\pm$ 0.55	74.29	<u>62.77$\pm$0.23</u>	<u>62.45</u>
	LFM	92.80 $\pm$ 0.05	<u>92.74</u>	76.15 $\pm$ 0.35	<u>75.51</u>	62.57 $\pm$ 0.34	61.88
	OPM	92.60 $\pm$ 0.22	92.21	75.17 $\pm$ 0.51	74.37	60.97 $\pm$ 0.78	59.66
	ACML	93.36$\pm$0.10	93.23	76.64$\pm$0.49	75.69	63.15$\pm$0.24	62.76

difference-aware attention constraint ($\mathcal{L}_{\text{DAC}}$). Furthermore, based on the final ACML framework, we further investigate the effectiveness of different optimization objectives and constraint formulations. Specifically, we replace $\mathcal{L}_{\text{DCTA}}$ with the standard supervised contrastive loss ($\mathcal{L}_{\text{SC}}$) and the confidence-constrained supervised contrastive loss ($\mathcal{L}_{\text{CSC}}$), respectively. Meanwhile, we replace the proposed $\mathcal{L}_{\text{DAC}}$ with a standard unweighted MSE constraint ($\mathcal{L}_{\text{MSE}}$) to analyze the impact of difference-aware weighting on the final performance. We report the average classification accuracy with standard deviation and the worst-case classification accuracy under different noise conditions.

Analysis. As shown in Table 2, introducing either $\mathcal{L}_{\text{DCTA}}$ or $\mathcal{L}_{\text{DAC}}$ consistently improves the performance compared with the original baseline model under different noise

Table 2: Ablation studies on NYU Depth V2. We report both the average performance and the worst-case performance under different noise levels.

Configuration	$\varepsilon = 0.0$		$\varepsilon = 5.0$		$\varepsilon = 10.0$	
	Avg $\pm$ Std	Worst	Avg $\pm$ Std	Worst	Avg $\pm$ Std	Worst
Baseline	70.87 $\pm$ 0.79	69.27	62.61 $\pm$ 1.53	58.87	55.02 $\pm$ 3.38	46.79
Baseline + $\mathcal{L}_{\text{DCTA}}$	72.37 $\pm$ 0.80	70.80	63.82 $\pm$ 1.97	58.56	55.70 $\pm$ 2.91	48.62
Baseline + $\mathcal{L}_{\text{DAC}}$	71.67 $\pm$ 0.82	70.32	63.66 $\pm$ 2.58	58.42	53.93 $\pm$ 4.33	47.59
ACML	72.72$\pm$0.71	71.56	65.43$\pm$1.63	60.86	58.11$\pm$2.69	49.69
ACML ($\mathcal{L}_{\text{DCTA}} \rightarrow \mathcal{L}_{\text{SC}}$)	72.48 $\pm$ 0.69	71.10	65.06 $\pm$ 1.74	60.86	57.09 $\pm$ 2.75	51.07
ACML ($\mathcal{L}_{\text{DCTA}} \rightarrow \mathcal{L}_{\text{CSC}}$)	72.55 $\pm$ 0.76	71.41	64.67 $\pm$ 1.55	60.09	57.43 $\pm$ 2.39	51.38
ACML ($\mathcal{L}_{\text{DAC}} \rightarrow \mathcal{L}_{\text{MSE}}$)	72.49 $\pm$ 0.71	71.10	64.78 $\pm$ 2.57	58.72	57.41 $\pm$ 3.15	49.69

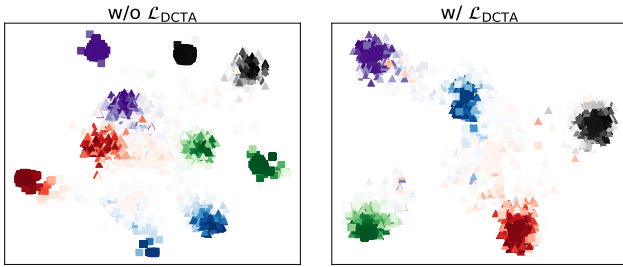


Figure 2: 2D t -SNE visualization of multimodal features. $\triangle$ indicates the image modality, and $\square$ denotes the text modality. Each color corresponds to a distinct class, with darker shades indicating higher confidence.

conditions, demonstrating their effectiveness in enhancing cross-modal representation learning. Moreover, the complete ACML framework achieves the best performance in terms of both average classification accuracy and worst-case classification accuracy, further validating the complementarity between $\mathcal{L}_{\text{DCTA}}$ and $\mathcal{L}_{\text{DAC}}$. Furthermore, replacing $\mathcal{L}_{\text{DCTA}}$ with either $\mathcal{L}_{\text{SC}}$ or $\mathcal{L}_{\text{CSC}}$ leads to inferior performance, demonstrating the superiority of the proposed dynamic cross-modal triplet alignment strategy in regularizing cross-modal representation distributions. Similarly, replacing $\mathcal{L}_{\text{DAC}}$ with the standard unweighted $\mathcal{L}_{\text{MSE}}$ constraint also degrades the performance, validating the effectiveness of the proposed difference-aware weighting mechanism.

4.3 RQ3 (Visualization Analysis)

Setting. To verify whether $\mathcal{L}_{\text{DCTA}}$ effectively reshapes the representation space and improves cross-modal semantic alignment, we perform t -SNE visualizations on the FOOD101 dataset to qualitatively examine the feature distributions before and after applying $\mathcal{L}_{\text{DCTA}}$.

Analysis. As shown in Figure 2, the t -SNE visualizations further reveal that, after introducing $\mathcal{L}_{\text{DCTA}}$, high-confidence samples form more compact clusters with improved cross-modal consistency. In contrast, low-confidence samples (indicated by lighter colors) remain relatively scattered, indicating that the proposed loss effectively promotes cross-modal alignment.

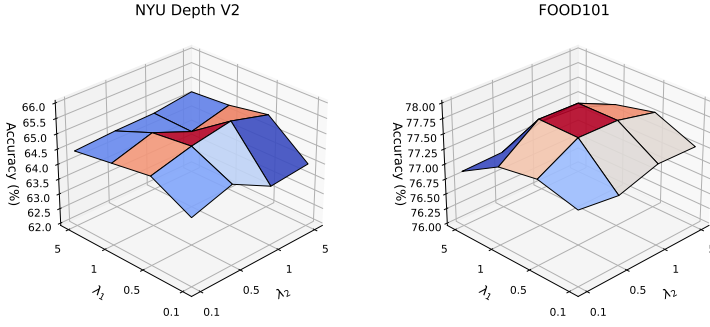
Figure 3: Performance sensitivity to hyperparameters λ_1 and λ_2 across two datasets.

Table 3: Accuracy comparison on FOOD101 validation and test sets under different missing rates (baselines vs. +ACML)

Missing rate η	Train/Test		Validation Set				Testing Set			
	Image	Text	DCP	(+ACML)	SyP	(+ACML)	DCP	(+ACML)	SyP	(+ACML)
50%	100%	50%	82.33	84.72	82.64	83.68	82.11	84.42	81.93	83.85
	50%	100%	89.23	89.44	89.07	89.16	89.12	89.44	89.18	89.39
	75%	75%	85.25	85.86	85.20	86.28	85.24	86.03	86.17	86.50
70%	100%	30%	79.18	80.61	79.39	80.21	78.87	80.39	79.28	80.27
	30%	100%	87.53	88.11	87.62	87.87	87.32	87.77	87.27	87.75
	65%	65%	82.38	83.24	82.53	83.55	81.87	83.11	82.25	83.24
90%	100%	10%	75.54	77.21	76.07	77.05	75.26	77.54	76.30	77.47
	10%	100%	86.26	86.36	86.50	86.59	85.78	86.00	86.19	86.55
	55%	55%	80.39	80.77	80.64	81.48	79.87	79.93	79.31	80.59

4.4 RQ4 (Hyper-parameter Analysis)

Setting. We evaluate the sensitivity of two key hyper-parameters, λ_1 and λ_2 , on the NYU Depth V2 and FOOD101 datasets, which control the constraint strengths of $\mathcal{L}_{\text{DCTA}}$ and $\mathcal{L}_{\text{DAC}}$, respectively. The search spaces for both hyper-parameters are set to $\{0.1, 0.5, 1, 5\}$. We report the mean classification accuracy under three settings, including the clean setting, $\varepsilon = 5$, and $\varepsilon = 10$ noisy settings, where the final result is obtained by averaging the accuracies across the three settings.

Analysis. As illustrated in Figure 3, excessively small or large values of either λ_1 or λ_2 lead to noticeable performance degradation. In contrast, when the hyper-parameters are within a reasonable range (e.g., 0.5 and 1), the proposed method achieves more stable and robust performance. Similar trends can be observed for both hyper-parameters, and the phenomenon becomes more evident on the larger-scale FOOD101 dataset. These results demonstrate that the proposed method maintains good stability and robustness within a reasonable hyper-parameter range.

4.5 RQ5 (Applicability to Missing Modalities)

Setting. To evaluate the adaptability of ACML under missing-modality conditions, we integrate it as a plug-and-play module into two prompt-tuning-based approaches, DCP[24] and SyP[13]. Experiments are conducted on the FOOD101 dataset. Following prior works[24,

Table 4: Classification results (%) on the BRCA and ROSMAP datasets.

Method	BRCA			ROSMAP		
	Accuracy	Weighted F_1	Macro F_1	Accuracy	F_1	AUC
MOGONet	82.9 $\pm$ 1.8	82.5 $\pm$ 1.7	77.4 $\pm$ 1.7	81.5 $\pm$ 2.3	82.1 $\pm$ 1.2	87.4 $\pm$ 1.2
MMDyn	87.7 $\pm$ 0.3	88.0 $\pm$ 0.5	84.5 $\pm$ 0.5	84.2 $\pm$ 1.3	84.6 $\pm$ 0.7	91.2 $\pm$ 0.7
MLCLNet	86.4 $\pm$ 1.6	87.8 $\pm$ 1.7	82.6 $\pm$ 1.8	84.4 $\pm$ 1.5	85.2 $\pm$ 1.5	89.3 $\pm$ 1.1
DMIB	86.0 $\pm$ 1.6	86.0 $\pm$ 0.8	81.6 $\pm$ 0.9	84.9 $\pm$ 1.8	85.3 $\pm$ 1.7	91.6 $\pm$ 0.7
GCFANet	88.6 $\pm$ 1.5	88.9 $\pm$ 1.6	85.3 $\pm$ 1.6	86.3 $\pm$ 1.4	88.3 $\pm$ 1.6	91.5 $\pm$ 1.2
HCMF	89.9 $\pm$ 0.2	90.1 $\pm$ 0.2	86.4 $\pm$ 0.6	88.0 $\pm$ 0.7	87.9$\pm$1.2	91.9 $\pm$ 0.7
ACML	90.7$\pm$0.2	91.1$\pm$0.2	88.1$\pm$0.2	88.2$\pm$0.8	<u>87.6$\pm$0.8</u>	92.4$\pm$0.7

[43], we consider three missing ratios: 50%, 70%, and 90%. For each ratio, we further evaluate three missing patterns: text missing, image missing, and both modalities missing, where each modality is removed at a ratio of 50%. All methods keep the backbone encoders frozen and only update the prompts together with our module. Performance is reported on both the validation and test sets.

Analysis. As reported in Table 3, ACML consistently improves the performance of both DCP and SyP across all missing-modality ratios and missing patterns on both validation and test sets. Notably, the gains remain stable even under severe modality absence (e.g., 90% missing ratio), demonstrating the strong robustness of ACML to incomplete multimodal inputs. These results suggest that ACML can effectively enhance representation learning when modality information is partially unavailable and can be seamlessly integrated into different prompt-tuning frameworks as a general and effective enhancement module.

4.6 RQ6 (Generalization and Scalability)

Setting. To evaluate the adaptability of ACML in broader multi-modal scenarios, we conduct classification experiments on two medical multi-omics benchmark datasets: BRCA[53] and ROSMAP[44, 45]. These tasks involve three sequential omics modalities ($M = 3$) and particularly emphasize population-level relationship modeling. Specifically, the BRCA dataset contains 875 samples spanning five subtypes, while the ROSMAP dataset includes 351 samples for binary Alzheimer’s disease diagnosis. For the network architecture, each modality employs a lightweight encoder—consisting of a single linear layer followed by a ReLU activation—to project input features into a 2048-dimensional space ($d_m = 2048$), with an MLP used for feature dimension alignment ($d = 2048$). The model is trained for 500 epochs using the Adam optimizer with a learning rate of 10^{-4} . For evaluation, the BRCA dataset uses accuracy, macro F_1 -score, and weighted F_1 -score, while the ROSMAP dataset adopts accuracy, F_1 -score, and AUC to comprehensively assess diagnostic performance. All results are reported as the mean over 20 random seeds.

Analysis. We compare ACML against six SOTA multimodal methods that focus on sequence-level modeling: MOGONet [53], MMDyn [47], MLCLNet [44], DMIB [45], GCFANet [48], and HCMF [8]. As shown in Table 4, ACML achieves significant and consistent performance gains across all tasks. These results demonstrate that our method maintains strong generalization capability across different data types and more diverse multi-modal scenarios.

5 Conclusion

This paper proposes an Alignment- and Calibration-driven Multimodal Learning framework (ACML) to address the limitations of existing methods in static alignment, which overlooks sample-wise differences and fusion strategies lacking difference-aware modeling. Specifically, ACML consists of two key modules: (i) a dynamic cross-modal triplet alignment (DCTA) module, which adaptively adjusts alignment strength according to sample confidence, enforcing cross-modal semantic consistency for high-confidence samples while preserving potential complementary information in low-confidence samples through an orthogonality encouragement strategy; and (ii) a difference-aware attention calibration (DAC) module, which adaptively adjusts the attention regularization strength according to feature magnitude differences and confidence differences across modalities, enabling more reasonable sample-aware fusion while avoiding unnecessary over-regularization. Experimental results demonstrate that ACML consistently achieves superior performance and robustness across multiple multimodal benchmark datasets, validating the effectiveness of the proposed modules. In future work, we plan to explore the integration of ACML with large-scale pre-trained foundation models, extending its applicability to complex real-world scenarios such as medical multimodal analysis, as well as multimodal understanding and generation tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62276052), and the Medico Engineering Cooperation Funds from University of Electronic Science and Technology of China and West China Hospital (No. ZYGX2022YGRH009).

References

- [1] David A. Bennett, Julie A. Schneider, Zoe Arvanitakis, and Robert S. Wilson. Overview and findings from the religious orders study. *Current Alzheimer Research*, 9 (6), 2012.
- [2] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16 (6), 2010.
- [3] Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In *ICML*, 2024.
- [4] Philip L De Jager, Yiyi Ma, Cristin McCabe, et al. A multi-omic atlas of the human frontal cortex for aging and alzheimer’s disease research. *Scientific Data*, 5(1), 2018.
- [5] Saisai Ding, Juncheng Li, Jun Wang, Shihui Ying, and Jun Shi. Multimodal co-attention fusion network with online data augmentation for cancer subtype classification. *IEEE Transactions on Medical Imaging*, 43(11), 2024.
- [6] Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *ICLR*, 2025.

- [7] Yingying Fang, Shuang Wu, Sheng Zhang, et al. Dynamic multimodal information bottleneck for multimodality classification. In *WACV*, 2024.
- [8] Yanglan Gan, Hangkai Zhao, Kaili Wang, Cairong Yan, and Guobing Zou. Hcmf: Hierarchical feature aggregation and cross-modal attention fusion framework for multi-omics patient classification. *IEEE Journal of Biomedical and Health Informatics*, 2025.
- [9] AV Geetha, T Mala, D Priyanka, and E Uma. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions. *Information Fusion*, 105, 2024.
- [10] Kangfu Han, Dan Hu, Fenqiang Zhao, et al. Incomplete multi-modal disentanglement learning with application to alzheimer’s disease diagnosis. *IEEE Transactions on Medical Imaging*, 2025.
- [11] Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *CVPR*, 2022.
- [12] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [13] Lingdong Kong, Xiang Xu, Jiawei Ren, et al. Multi-modal data-efficient 3d scene understanding for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [14] Deng Li, Bohao Xing, Xin Liu, et al. Deemo: De-identity multimodal emotion recognition and reasoning. In *ACM MM*, 2025.
- [15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.
- [16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.
- [17] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey: S. li, h. tang. *International Journal of Computer Vision*, 134(3), 2026.
- [18] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *NeurIPS*, 35, 2022.
- [19] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2), 2021.
- [20] Huan Ma, Zongbo Han, Changqing Zhang, et al. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. *NeurIPS*, 34, 2021.

- [21] Qianxia Ma, Ming Zhang, Yan Tang, and Zhen Huang. Att-sinkhorn: Multimodal alignment with sinkhorn-based deep attention architecture. In *ICAC*, 2023.
- [22] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *NeurIPS*, 34, 2021.
- [23] Jiahui Qu, Yuanbo Yang, Wenqian Dong, and Yufei Yang. Lds2ae: Local diffusion shared-specific autoencoder for multimodal remote sensing image classification with arbitrary missing modalities. In *AAAI*, volume 38, 2024.
- [24] Tongkai Shi, Wei Feng, Fanhua Shang, Liang Wan, et al. Deep correlated prompting for visual recognition with missing modalities. *NeurIPS*, 37, 2024.
- [25] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.
- [26] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015.
- [27] Xiaorui Su, Pengwei Hu, Dongxu Li, et al. Interpretable identification of cancer genes across biological networks via transformer-powered graph representation learning. *Nature Biomedical Engineering*, 2025.
- [28] Xueling Suo, Mengyao Chen, Li Chen, et al. Automatic identification of parkinsonism using clinical multi-contrast brain mri: a large self-supervised vision foundation model strategy. *EBioMedicine*, 116, 2025.
- [29] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *CVPR*, 2023.
- [30] Qun Wang, Feng Zhu, Ge Wu, et al. Object-level and scene-level feature aggregation with clip for scene recognition. *Information Fusion*, 120, 2025.
- [31] Tongxin Wang, Wei Shao, Zhi Huang, et al. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1), 2021.
- [32] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *ICMEW*, 2015.
- [33] Zheng Wang, Zhenwei Gao, Yang Yang, et al. Geometric matching for cross-modal retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3), 2024.
- [34] Yake Wei, Di Hu, Henghui Du, and Ji-Rong Wen. On-the-fly modulation for balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1), 2025.
- [35] Jie Xu, Huayi Tang, Yazhou Ren, et al. Multi-level feature learning for contrastive multi-view clustering. In *CVPR*, 2022.
- [36] Chonghua Xue, Sahana S Kowshik, Diala Lteif, et al. Ai-based differential diagnosis of dementia etiologies on multimodal data. *Nature Medicine*, 30(10), 2024.

- [37] Zihui Xue and Radu Marculescu. Dynamic multimodal fusion. In *CVPR*, 2023.
- [38] Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. *NeurIPS*, 37, 2024.
- [39] Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. Multimodal contrastive training for visual representation learning. In *CVPR*, 2021.
- [40] Ying Zeng, Wenjun Yan, Sijie Mai, and Haifeng Hu. Disentanglement translation network for multimodal sentiment analysis. *Information Fusion*, 102, 2024.
- [41] Qingyang Zhang, Haitao Wu, Changqing Zhang, et al. Provable dynamic fusion for low-quality multimodal data. In *ICML*, 2023.
- [42] Qingyang Zhang, Yake Wei, Zongbo Han, et al. Multimodal fusion on low-quality data: A comprehensive survey. *Information Fusion*, 135, 2026.
- [43] Zhihui Zhang, Luanyuan Dai, Qika Lin, Yunfeng Diao, Guangyin Jin, Yufei Guo, Jing Zhang, and Xiaoshuai Hao. Synergistic prompting for robust visual recognition with missing modalities. In *ICCV*, 2025.
- [44] Xiao Zheng, Chang Tang, Zhiguo Wan, Chengyu Hu, and Wei Zhang. Multi-level confidence learning for trustworthy multimodal classification. In *AAAI*, volume 37, 2023.
- [45] Xiao Zheng, Minhui Wang, Kai Huang, and En Zhu. Global and cross-modal feature aggregation for multi-omics data classification and application on drug response prediction. *Information Fusion*, 102, 2024.
- [46] Wujie Zhou, Shaohua Dong, Jingsheng Lei, and Lu Yu. Mtanet: Multitask-aware network with hierarchical multimodal fusion for rgb-t urban scene understanding. *IEEE Transactions on Intelligent Vehicles*, 8(1), 2022.
- [47] Xiaoguang Zhu, Lianlong Sun, Yang Liu, et al. Causal debiasing medical multimodal representation learning with missing modalities. *arXiv preprint arXiv:2509.05615*, 2025.