ISSN 2616-5775 Vol. 8, Issue 7: 87-94, DOI: 10.25236/AJCIS.2025.080711

Cross-Modal Adaptive Fusion and Enhancement Model for Noise-Robust Scenarios

Wenhui Zhang[1,a,*], Qianxi Li[1,b]

[1]*School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing, 400074, China*
[a]*17385339798@163.com*, [b]*lqxdyx0121@163.com*
**Corresponding author*

Abstract: *With the rapid advancement of computer technology, multi-modality has emerged as a critical area of research. The fusion and alignment of multi-modal data not only enhance the intelligence level of Internet of Things (IoT) devices but also provide users with enriched and precise service experiences. However, most existing studies primarily focus on managing two to three modalities, which often proves inadequate in real-world complex and dynamic scenarios. To address this limitation, this paper conducts an in-depth investigation into multi-modal learning with the aim of overcoming the constraints associated with current modality quantities. In practical applications, coordinating multiple modalities remains a significant challenge, particularly in dynamic environments where noise factors can lead to fluctuating modality dominance. Consequently, achieving effective multi-modal fusion and alignment has become a key research challenge. This paper proposes a novel multi-modal fusion framework that emphasizes both inter-modal complementarity and collaboration while introducing a modality enhancement mechanism designed to mitigate noise interference across modalities. Experimental results validate the effectiveness of our proposed method across four benchmark datasets.*

Keywords: *Multi-Modal Learning, Cross-Modal Alignment, Modality Fusion, Contrastive Learning, Noise Robustness*

1. Introduction

In recent years, with ongoing technological advancements, multi-modality approaches have been extensively applied across various computational tasks and have increasingly infiltrated key domains such as the Internet of Things (IoT)[1]. As IoT technologies evolve rapidly, there is an escalating demand for intelligent services driven by diverse sensor data. In this context, multi-modality learning has emerged as a pivotal area of research. Its primary objective is to fully leverage the complementary information among different modalities to enhance the perceptual capabilities as well as cognitive and decision-making processes within IoT systems. The fusion and alignment of multi-modality data can significantly elevate the intelligence level of IoT devices while providing users with richer and more accurate service experiences in intricate environments[2].

As a result, multi-modality fusion and modality alignment have increasingly become central techniques in contemporary multi-modality research. Modality alignment primarily addresses semantic discrepancies and distributional inconsistencies among various data types. Once alignment is achieved, modality fusion further integrates multi-modality data—originating from images, speech, text, and sensors—through methods such as weighted combination, attention mechanisms, and feature transformation. This process generates joint representations characterized by strong expressiveness and discriminative power[3-5]. Consequently, it enhances overall data utilization efficiency while significantly broadening the application boundaries of multi-modality systems within the Internet of Things (IoT).

However, most existing multi-modality research remains limited to relatively simple application scenarios, typically involving only two or three modalities[6-7]. Such constraints often result in poor adaptability and scalability when confronted with the complex and dynamic multi-modality environments of the real world. Moreover, in dynamic settings, the quality and contribution of each modality can fluctuate due to factors such as noise, occlusion, signal attenuation, or modality failure. These variations can lead to frequent shifts in the relative strengths of different modalities, thereby posing significant challenges for achieving stable fusion.To address these challenges, this paper proposes a novel multi-modality fusion approach that not only emphasizes complementary collaboration among

ISSN 2616-5775 Vol. 8, Issue 7: 87-94, DOI: 10.25236/AJCIS.2025.080711

modalities during the fusion process but also fully accounts for the dynamic variation in modality performance within practical applications.

2. Related Works

2.1 Modal Alignment and Fusion

Modality alignment is a crucial aspect of multi-modality machine learning, as it establishes correspondences between data from different modalities to facilitate effective information complementation. By associating semantic or structural information across modalities, modality alignment enables the model to infer information from one modality based on another, thereby alleviating distributional discrepancies between modalities.Early studies predominantly relied on supervised alignment techniques. For instance, DeVISE maximizes the similarity between image features and their corresponding textual labels within a shared embedding space to achieve supervised alignment[8]. In recent years, self-supervised alignment learning has emerged as one of the prevailing approaches in this domain. CLIP employs large-scale contrastive training with image-text pairs to project both modalities into proximate positions within a joint embedding space while simultaneously distancing unrelated instances[9]. This methodology significantly enhances zero-shot cross-modal retrieval capabilities. Similarly, ALIGN capitalizes on naturally occurring alignments between images and texts found in noisy web data, further illustrating the effectiveness of large-scale alignment learning[10]. However, many depend on well-aligned multi-modality datasets when in practice cross-modal datasets often contain noise and weakly aligned pairs.To address these issues, Xiao X proposed a more refined approach by identifying tokens in text that are highly relevant to their corresponding images and assigning them greater weight within the loss function. This strategy improves alignment accuracy by concentrating the learning process on more informative multi-modality correspondences[11].

Modality fusion emphasizes the integration of complementary information from multiple modalities, typically categorized into early fusion, late fusion, and hybrid fusion. Early fusion, often referred to as feature-level fusion, is one of the most widely utilized strategies in multi-modality learning. This approach involves merging features extracted from each modality prior to executing downstream analytical tasks. Current methodologies for feature-level fusion predominantly encompass probabilistic statistical models, neural network-based techniques, feature extraction methods, and search-based strategies[12]. In contrast, late fusion entails the independent processing of data and features from different modalities through separate models; each model generates unimodal decisions. Hybrid fusion integrates both feature-level and decision-level approaches with the objective of harnessing the advantages of early and late fusion while addressing their respective limitations. For example, ViLBERT employs cross-modal attention mechanisms that facilitate interaction between visual and textual features across Transformer layers[13]. Building on this foundation, LXMERT introduces a task-driven gating mechanism designed to adaptively select salient modality-specific features, thereby enhancing the model's ability to concentrate on task-relevant information[14].

2.2 Modal Enhancement

Due to real-world challenges such as modality-specific noise, missing modalities, and low-quality inputs, it is essential to enhance unimodal representations and their robustness in order to establish a more reliable foundation for cross-modal alignment and fusion. Current research on modality enhancement can be broadly categorized into two primary domains:

The first category pertains to unimodal enhancement. Early investigations concentrated on improving single-modality representations through techniques such as data augmentation or feature optimization. For instance, SpecAugment introduces random masking within the time-frequency domain of speech signals, significantly bolstering the noise robustness of speech recognition models[15]. In the visual domain, RandAugment employs automated search methods to identify optimal combinations of augmentation strategies, thereby addressing the limitations associated with traditional handcrafted approaches[16]. While these techniques enhance the generalization capabilities of unimodal models, they often overlook semantic consistency across modalities; this oversight may lead to misaligned features that diverge from the objectives inherent in cross-modal tasks.

The second category pertains to cross-modal enhancement, wherein researchers utilize information from various modalities to enhance unimodal representations, particularly in light of the emergence of large-scale multi-modality pre-trained models. For example, CMKD introduces a cross-modal

ISSN 2616-5775 Vol. 8, Issue 7: 87-94, DOI: 10.25236/AJCIS.2025.080711

knowledge distillation framework that facilitates the transfer of alignment knowledge from a multi-modality teacher model to a unimodal student network[17]. This approach enables the student network to preserve discriminative features even when certain modalities are absent. Similarly, AV-HuBERT employs audio-visual contrastive learning to constrain the latent space associated with lip movements, thereby indirectly bolstering the robustness of unimodal lip-reading encoders[18]. However, these methods heavily rely on the quality of cross-modal alignment and may introduce noise in scenarios characterized by weak inter-modality correlations.

In summary, current mainstream approaches are transitioning from traditional static fusion strategies towards modality-adaptive fusion techniques. Inspired by this trend, we propose a comprehensive multi-modality fusion framework that integrates a contrastive loss-based modality enhancement mechanism. This design further enhances feature modeling for each modality during practical tasks, thus improving fusion robustness under challenging conditions such as noise.

3. Proposed method

Figure 1 presents the overall architecture of our proposed model. Initially, each modality is converted into a unified sequence format, Subsequently, modality-specific features are extracted utilizing a multi-modality encoder. After computing the weak modality through cosine similarity, we employ contrastive learning to enhance its representation.For modality fusion, a cross-attention mechanism is utilized to integrate information across different modalities. Detailed descriptions of these processes can be found in Sections 3.1 and Sections 3.2.

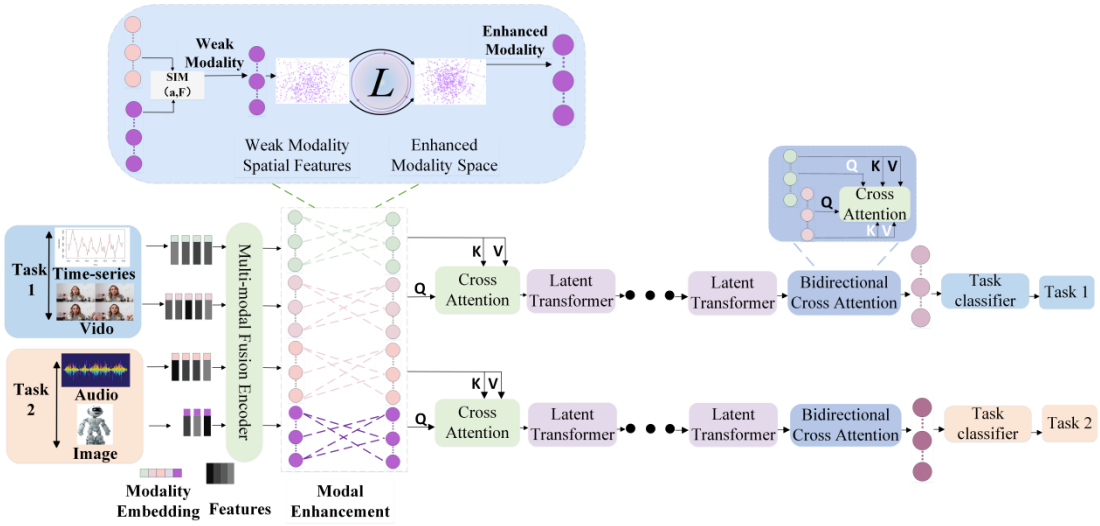

Figure 1: The model architecture diagram.

3.1 Modality Fusion and Decision Making

Since this paper involves many different modalities, we set the modality as $X_m \in \mathbb{R}$, where $m = \{1,2,\cdots,n\}$. First, we normalize each modality into a sequence format, meaning that each element in a table or figure is treated as an individual token in the sequence. As a result, we obtain the standardized input data $X_m = t_m \times d_m$ for subsequent processing, where t_m is the input sequence length specifically designed for the mode and task, and d_m is the input dimension specifically designed for the mode and task. For each different model $m \in M$, Liang and Paul Pu define a one-hot embedding mode $e_m \in \mathbb{R}^{|M|}$, where $|M|$ is the total number of modalities[19]. This embedding identifies the shared modality pattern across different tasks. Additionally, we introduce a modality-specific positional encoding $P_m \in \mathbb{R}^{t_m \times d_{p_m}}$, where d_{p_m} is the dimension of the positional encoding, used to capture the position of each element in the modality sequence.Moreover, we employ a shared positional encoding to capture temporal dimensions that are common across modalities.After processing, the final representation for modality m is: $X_m = X_m \oplus e_m \oplus P_m \oplus 0_m$. Here, $\oplus$ denotes concatenation of: the input sequence,modality embedding, positional encoding, and padding zeros. The padding 0_m is used to fill the sequence with zeros so that all modalities are padded to the same standardized length $t_m \times d_{all}$, where $d_{all} = \max_{m \in M}(d_m + |M| + d_{p_m})$ and d_{all} is the final channel dimension of the modality representation.

ISSN 2616-5775 Vol. 8, Issue 7: 87-94, DOI: 10.25236/AJCIS.2025.080711

Based on this unified representation of all modalities, we adopt it as the input format and design a general-purpose encoder based on the Transformer Perceiver architecture proposed by Andrew Jaegle[20]:

$$\begin{aligned} \overline{Z}_m^{(L)} &= Cross \quad Attention\left(Z_m^{L-1}, X_m\right) \\ &= soft\max\left(\frac{Q_C K_c^{\mathrm{T}}}{\sqrt{d_{LS}}}\right) V_c \\ &= soft\max\left(\frac{Z_m^{(L-1)} W_{Q_c} W_{V_c}^{\mathrm{T}} X_m^{\mathrm{T}}}{\sqrt{d_{LS}}}\right) X_m W_{V_c} \end{aligned} \tag{1}$$

The model is trained recursively with a latent array $\mathrm{Z_m}$ of shape $\mathrm{d_{LN}} \times \mathrm{d_{LS}}$, where $\mathrm{d_{LN}}$ denotes the sequence length of the latent dimension, and $\mathrm{d_{LS}}$ represents the latent feature dimension. The latent array is randomly initialized as $\mathrm{Z_m^{(0)}}$ at the beginning. Before computing each layer, we require the latent array from the previous layer $\mathrm{Z_m^{(L-1)}}$ to calculate the current layer. The processed input $\mathrm{X_m}$ interacts with $\mathrm{Z_m^{(L-1)}}$ through cross-attention to produce an intermediate representation $\mathrm{\bar{Z}_m^L}$. Then, self-attention is applied to $\mathrm{\bar{Z}_m^L}$ to obtain the input representation $\mathrm{Z_m^{(L)}}$ for the next layer. Repeating the above steps iteratively yields the final latent representation $\mathrm{Z_m^{(L)}} \in \mathbb{R}^{\mathrm{d_{LN}} \times \mathrm{d_{LS}}}$. Let $\mathrm{W_{Q_c}} \in \mathbb{R}^{\mathrm{d_{LS}} \times \mathrm{d_{LS}}}$, $\mathrm{W_{K_c}} \in \mathbb{R}^{\mathrm{d_{all}} \times \mathrm{d_{LS}}}$, $\mathrm{W_{V_c}} \in \mathbb{R}^{\mathrm{d_{all}} \times \mathrm{d_{LS}}}$ be the learnable parameters used in cross-attention, while $\mathrm{W_{Q_s}} \in \mathbb{R}^{\mathrm{d_{LS}} \times \mathrm{d_{LS}}}$, $\mathrm{W_{K_s}} \in \mathbb{R}^{\mathrm{d_{LS}} \times \mathrm{d_{LS}}}$, $\mathrm{W_{V_s}} \in \mathbb{R}^{\mathrm{d_{LS}} \times \mathrm{d_{LS}}}$ are parameters for self-attention.

$$\begin{aligned} Z_m^{(L)} &= Self \quad Attention\left(\overline{Z}_m^{(L)}\right) \\ &= soft\max\left(\frac{Q_s K_s^{\mathrm{T}}}{\sqrt{d_{LS}}}\right) V_s \\ &= soft\max\left(\frac{\overline{Z}_m^{(L)} W_{Q_s} W_{V_s}^{\mathrm{T}} \overline{Z}_m^{(L)\mathrm{T}}}{\sqrt{d_{LS}}}\right) \overline{Z}_m^{(L)} W_{V_s} \end{aligned} \tag{2}$$

To effectively model multi-modality representations, we adopt a shared cross-modal Transformer block with parameter C, as proposed by Yao-Hung and Jiasen Lu[21-22]. Within this framework, for any pair of latent arrays $\mathrm{Z_1}$and $\mathrm{Z_2}$from two individual modalities, we compute the attention weights in a cross-modal manner using the shared Transformer block with parameter C. Specifically, we designate $\mathrm{Z_1}$ as the query input and $\mathrm{Z_2}$ as both the key and value, forming the basis of the attention mechanism to model the interactions from $\mathrm{Z_1}$ to $\mathrm{Z_2}$:

$$\begin{aligned} Z_{1\to 2} &= Cross \quad Attention\left(Z_2, Z_1\right) \\ &= soft\max\left(\frac{Q_2 K_1^{\mathrm{T}}}{\sqrt{d_k}}\right) V_1 \\ &= soft\max\left(\frac{Z_2 W_{Q_2} W_{V_1}^{\mathrm{T}} Z_1^{\mathrm{T}}}{\sqrt{d_k}}\right) Z_1 W_{V_1} \end{aligned} \tag{3}$$

Here, $W_{Q_2} \in \mathbb{R}^{d_{LS} \times d_k}$, $W_{V_1} \in \mathbb{R}^{d_{LS} \times d_k}$ are learnable parameters. Based on the equations (3), in the same way, we can also get attention between: $Z_{2\to 1}$, the final multi-modality representation is formed by concatenating the bidirectional attention results:

$$Z_{mm} = \left[Z_{1\to 2}, Z_{2\to 1}\right] \tag{4}$$

To determine the relative strength or weakness of each modality, we employ residual similarity as a basis for comparison. After acquiring the representations of each modality, we first perform feature-level fusion to construct a shared reference embedding. Specifically, the fused representation is obtained as:

$$z_{fused} = Concat\left(z_1, z_2, \ldots, z_i\right) \tag{5}$$

Next, we measure the cosine similarity between the individual modality representation $\mathrm{Z_i}$ and the fused feature vector $\mathrm{Z_{fused}}$ to assess alignment. The average similarity score across the dataset is computed as:

ISSN 2616-5775 Vol. 8, Issue 7: 87-94, DOI: 10.25236/AJCIS.2025.080711

$$\mathrm{sim}_{\mathrm{i}}^{(\mathrm{n})} = \sum_{n=1}^{N} \frac{z_i^{(n)} \cdot z_{fused}^{(n)}}{\| z_i^{(n)} \| \cdot \| z_{fused}^{(n)} \|} \tag{6}$$

Where N denotes the number of samples, and $\mathrm{sim}_{\mathrm{i}}^{(\mathrm{n})}$ represents the average similarity between mo$\tilde{\mathrm{Z}}_1$dality i and the fused vector across all instances. Finally, the modality with the lowest similarity score is identified as the weak modality:

$$Weak \quad Modality = \arg\min_i \ sim_i^{(n)} \tag{7}$$

3.2 Enhancement of Weak Modalities

The contrastive loss function serves as a key component for model optimization. It enhances the semantic alignment between weak modality representations and the fused embedding, thereby improving the consistency and effectiveness of multi-modality fusion. To this end, we consider two modalities $\mathrm{Z}_1$and Z_2 and construct a batch of sample representations. In this subsection, we describe the procedures of vector normalization, similarity computation, and similarity matrix construction, which are essential for the enhancement process.

Vector Normalization: To mitigate the influence of scale on similarity calculation, we first apply $\mathcal{L}_2$ normalization to both vectors $\mathrm{Z}_1$and Z_2. Among them, Z_1 is weak modality and Z_2 is strong modality:

$$\tilde{Z}_1 = \frac{Z_1}{\| Z_1 \|} \qquad \tilde{Z}_2 = \frac{Z_2}{\| Z_2 \|} \tag{8}$$

After normalization, vectors are projected onto a unit hypersphere, and the residual similarity between them reduces to their dot product, i.e.: $\tilde{\mathrm{Z}}_1 \cdot \tilde{\mathrm{Z}}_2 = \cos\theta$, which reflects the angular similarity. This operation ensures that similarity measurements focus purely on direction rather than magnitude, making the comparison more robust.

Residual Similarity Matrix Construction: For a batch of B sample pairs, we compute pairwise similarities between the normalized representations of $\mathrm{Z}_1$and Z_2, yielding a residual similarity matrix $\mathrm{S} \in \mathbb{R}^{\mathrm{B}\times\mathrm{B}}$, where:

$$S_{i,j} = \tilde{Z}_{1,i} \cdot \tilde{Z}_{2,j} \tag{9}$$

Here,$\mathrm{S}_{\mathrm{i,j}}$ represents the residual similarity between the i-th sample in Z_1 and the j-th sample in Z_2. The matrix captures how each sample aligns with others across modalities (e.g., image-text or multi-view representations), providing a foundation for identifying and enhancing weak modalities.

Contrastive Loss Design: To distinguish between positive and negative pairs, we adopt a temperature-scaled Softmax function for normalization. For the i-th sample, the contrastive loss is defined as:

$$L_{contrast} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{e^{S_{i,i}/\tau}}{\sum_{j=1}^{B}\left(e^{S_{i,j}/\tau} + e^{S_{j,i}/\tau}\right) - e^{S_{i,i}/\tau}} \tag{10}$$

Here, the temperature parameter τ controls the sharpness of the similarity distribution. A lower τ increases the model's focus on distinguishing hard negative samples by sharpening the Softmax distribution. Conversely, higher values of τ lead to smoother distributions and reduced training difficulty. To avoid redundancy from symmetric angular distances, the denominator includes both $\mathrm{S}_{\mathrm{i,j}}$ and $\mathrm{S}_{\mathrm{j,i}}$, excluding the self-similarity term. This design follows the symmetric contrastive learning framework proposed in Chen T, which aims to emphasize positive sample similarity while suppressing the influence of hard negatives[23].

Regularization Term: To prevent the model from collapsing due to over-reliance on similarity optimization, we introduce an embedding norm regularization term $\mathcal{L}_{\mathrm{reg}}$, defined as:

$$L_{reg} = \lambda\left(\| Z_1 \|_2^2 + \| Z_2 \|_2^2\right) \tag{11}$$

This term penalizes excessively large feature norms, encouraging a balance between representation learning and the primary task. The hyperparameter λ controls the strength of the regularization.

ISSN 2616-5775 Vol. 8, Issue 7: 87-94, DOI: 10.25236/AJCIS.2025.080711

Final Loss Function: The overall objective is a weighted sum of the primary task loss, the contrastive loss, and the regularization term:

$$L = L_{task} + \alpha L_{contrast} + L_{reg} \tag{12}$$

Here, α is a hyperparameter that adjusts the importance of the contrastive loss. By balancing the contrastive supervision with the main task objectives, the model dynamically enhances weak modality representations and improves the robustness and integrity of the multi-modality system.

This work proposes a comprehensive modality fusion and enhancement framework, which integrates multiple modalities, determines modality strength through cosine similarity, and applies contrastive learning to enhance weaker modalities. The proposed approach not only extends fusion to support multiple modalities but also effectively improves the adaptability of the multi-modality fusion system to varying modality conditions, thereby enhancing the overall robustness and generalization capability of the system.

4. Experimental Analysis

4.1 Dataset

In this section, we present comparative experiments aimed at evaluating the effectiveness of our proposed model. The experiments are conducted on a diverse array of multi-modality datasets provided by MultiBench, encompassing six modalities. Specifically, the datasets utilized include:

UR-FUNNY: This dataset comprises 16,514 samples and incorporates three modalities—text, video, and audio—focusing on humor recognition.

MOSEI: Similar to UR-FUNNY in terms of modalities, MOSEI is a larger dataset containing 22,777 samples and is designed for sentiment analysis utilizing text, visual, and acoustic inputs.

MIMIC: Comprising 36,212 instances, this clinical dataset includes time-series signals as well as tabular data to support medical diagnosis and prediction tasks.

AV-MNIST: This dataset consists of 70,000 samples that each combine image and audio modalities; it serves as a benchmark for audio-visual classification.

4.2 Experimental Environment

This paper conducts experiments on the pytorch framework under Ubuntu 22.04, using two RTX 3090 graphics cards and 120G of memory. The Adam optimization model is employed, with a learning rate of 5×10^{-4}. In the model architecture design, we set the number of latent periods to 20, the latent dimension to 64, and the L_2 regularization to 3×10^{-4}. Additionally, the model runs for 100 epochs, which takes approximately 168 hours.

4.3 Evaluation and Analysis

In this subsection, we assess the performance of the proposed method across four benchmark datasets: UR-FUNNY, MOSEI, MIMIC, and AV-MNIST, utilizing Accuracy (Acc) as the primary evaluation metric. We conduct comparative experiments and provide a comprehensive analysis and discussion of the results to validate the effectiveness and reliability of our approach. The comparative results are presented in Table 1. In this table:

MULTIBENCH serves as a comprehensive and unified benchmark for multi-modality learning, encompassing 15 datasets, 10 modalities, 20 prediction tasks, and 6 research domains[24] .

VATT is a convolution-free Transformer model designed to directly process raw video, audio, and text inputs[25]. It employs a multi-modality contrastive learning strategy within a self-supervised learning framework.

HIGHMMT is a scalable multi-modality fusion model that leverages shared-parameter multi-task learning to enhance cross-modal and cross-task generalization through transfer learning mechanisms[19].

These models serve as robust baselines for comparison purposes while underscoring the competitiveness of our method across various multi-modality scenarios.

ISSN 2616-5775 Vol. 8, Issue 7: 87-94, DOI: 10.25236/AJCIS.2025.080711

Table 1: Comparative Experimental Results.

Model	MOSEI↑ Acc(%)	MIMIC↑ Acc(%)	AV-MNIST↑Acc(%)	UR-FUNNY↑Acc(%)	Ave↑ (%)
MULTIBENCH[24]	79.4	67.7	70.4	63.7	70.2
VATT[25]	79.5	64.3	70.1	63.0	69.2
HIGHMMT[19]	79.4	68.9	70.2	64.2	70.8
Our	80.2	68.3	70.2	66.4	71.3

As demonstrated in Table 1, our proposed method achieves superior performance on the MOSEI and UR-FUNNY datasets, attaining an accuracy of 80.2%. While the highest scores on the MIMIC and AV-MNIST datasets are achieved by HIGHMMT and MULTIBENCH, respectively, our approach exhibits strong competitiveness—showing only a 0.6% gap from the best-performing model on MIMIC and a mere 0.2% difference on AV-MNIST. Overall, when averaged across all datasets, our method consistently outperforms other approaches, achieving a 0.5% improvement over HIGHMMT, which serves as the strongest baseline among those compared.

The proposed multi-modality fusion and enhancement framework demonstrates strong overall performance across four benchmark datasets, with particularly notable results on UR-FUNNY and MOSEI, where it achieves the best accuracy—improving by 2.4% and 0.8%, respectively. Moreover, the model attains an average accuracy of 71.3%, outperforming the strongest baseline, HIGHMMT (70.8%).

The performance gains observed on UR-FUNNY and MOSEI indicate that the proposed modality enhancement mechanism effectively addresses challenges such as noise and dynamic variation across modalities. By applying contrastive learning to reinforce weaker modalities, the model suppresses the influence of poor-quality signals while strengthening high-contributing modalities.

5. Conclusions

To tackle challenges such as noise interference and variability in modality strength within multi-modality data fusion and alignment, this paper presents an enhanced framework grounded in cross-modal alignment and adaptive fusion. By integrating a modality enhancement mechanism alongside a contrastive learning strategy, the model effectively mitigates the impact of low-quality modalities while dynamically adjusting the fusion weights across different modalities.

References

[1] Yuan Y, Li Z, Zhao B. A Survey of Multimodal Learning: Methods, Applications, and Future[J]. ACM Computing Surveys, 2025.DOI:10.1145/3713070.
[2] Baltrusaitis T, Ahuja C, Morency L P. Multimodal Machine Learning: A Survey and Taxonomy[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99):1-1.DOI:10. 1109/TPAMI. 2018. 2798607.
[3] Qiao Y, Zhou H, Liu Y, et al. A multi-modal fusion model with enhanced feature representation for chronic kidney disease progression prediction[J]. Briefings in Bioinformatics, 2024, 26(1). DOI:10.1093/bib/bbaf003.
[4] Jiang J. Research on Ship-Type Recognition Based on Fusion of Ship Trajectory Image and AIS Time Series Data[J].Electronics, 2025, 14.DOI:10.3390/electronics14030431.
[5] Xue Z, Marculescu R. Dynamic Multimodal Fusion[J]. Computer Vision and Pattern Recognition Workshops, 2022.DOI:10.48550/arXiv.2204.00102.
[6] Chen C F, Fan Q, Panda R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification[J]. International Conference on Computer Vision, 2021.DOI:10.48550/arXiv.2103.14899.
[7] Park G Y , Kim J , Kim B ,et al.Energy-Based Cross Attention for Bayesian Context Update in Text-to-Image Diffusion Models[J]. Advances in Neural Information Processing Systems, 2023.
[8] Cao Y, Long M, Wang J, et al. Deep Visual-Semantic Quantization for Efficient Image Retrieval[C] //Computer Vision & Pattern Recognition.IEEE, 2017.DOI:10.1109/CVPR.2017.104.
[9] Radford A, Kim J W, Hallacy C, et al. Learning Transferable Visual Models From NaturalLanguage Supervision[J]. International Conference on Machine Learning, 2021.DOI:10.48550/arXiv.2103.00020.
[10] Jia C, Yang Y, Xia Y, et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision[J]. International Conference on Machine Learning, 2021. DOI:10.48550/arXiv.2102.05918.
[11] Xiao X, Wu B, Wang J, et al. Seeing the Image: Prioritizing Visual Correlation by Contrastive Alignment[J]. Computer Vision and Pattern Recognition (2024).

ISSN 2616-5775 Vol. 8, Issue 7: 87-94, DOI: 10.25236/AJCIS.2025.080711

[12] Xuelong Li. Multimodal Cognitive Computing [J]. Chinese Journal of Science: Information Science, 2023,53(01):1-32.
[13] Lu J, Batra D, Parikh D, et al. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks[C]//Advances in Neural Information Processing Systems 32, Volume 1 of 20: 32nd Conference on Neural Information Processing Systems (NeurIPS 2019).Vancouver(CA).8-14 December 2019.2020.
[14] Tan H, Bansal M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers [J]. Conference on Empirical Methods in Natural Language Processing, 2019. DOI:10.18653/v1/D19-1514.
[15] Park D S , Chan W , Zhang Y ,et al.SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition[J]. 2019.DOI:10.21437/Interspeech.2019-2680.
[16] Cubuk E D , Zoph B , Shlens J ,et al.Randaugment: Practical automated data augmentation with a reduced search space[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).IEEE, 2020.DOI:10.1109/CVPRW50498.2020.00359.
[17] Wei Z M, Pan H Y, Qiao L B, et al. Cross-modal knowledge distillation in multi-modal fake news detection[C]// Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 4733-4737. DOI: 10.1109/ICASSP43922.2022.9747280.
[18] Shi B , Hsu W N , Lakhotia K ,et al.Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction[J]. In Proc. International Conference on Learning Representations, 2022. DOI:10.48550/arXiv.2201.02184.
[19] Liang P P, Lyu Y, Fan X, et al. HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning[J]. Transactions on Machine Learning Research, 2022. DOI:10.48550/arXiv.2203.01311.
[20] Jaegle A, Gimeno F, Brock A, et al. Perceiver: General Perception with Iterative Attention[J]. 2021. DOI:10.48550/arXiv.2103.03206.
[21] Tsai Y H H, Bai S, Liang P P, et al. Multimodal Transformer for Unaligned Multimodal Language Sequences[J]. Proceedings of the conference. Association for Computational Linguistics. Meeting, 2019. DOI:10.18653/v1/P19-1656.
[22] Lu J, Batra D, Parikh D, et al. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks[C]//Advances in Neural Information Processing Systems 32, Volume 1 of 20: 32nd Conference on Neural Information Processing Systems (NeurIPS 2019).Vancouver(CA).8-14 December 2019.2020.
[23] Chen T, Kornblith S, Norouzi M, et al. A Simple Framework for Contrastive Learning of Visual Representations[J]. Machine Learning. 2020.DOI:10.48550/arXiv.2002.05709.
[24] Liang P P, Lyu Y, Fan X, et al. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning[J]. Machine Learning. 2021.DOI:10.48550/arXiv.2107.07502.
[25] Akbari H, Yuan L, Qian R, et al. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text[J]. Computer Vision and Pattern Recognition. 2021.DOI: 10. 48550/arXiv. 2104.11178.