

# Cross-modal Prompting for Balanced Incomplete Multi-modal Emotion Recognition

Wen-Jue He<sup>1</sup>, Xiaofeng Zhu<sup>2</sup>, Zheng Zhang<sup>1, 3\*</sup>

<sup>1</sup>Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen, China

<sup>2</sup>School of Computer Science and Technology, Hainan University, Haikou, China

<sup>3</sup>Shenzhen Loop Area Institute, Shenzhen, China

he\_wenjue@163.com, seanzhuxf@gmail.com, darrenzz219@gmail.com

## Abstract

Incomplete multi-modal emotion recognition (IMER) aims at understanding human intentions and sentiments by comprehensively exploring the partially-observed multi-source data. Although the multi-modal data is expected to provide more abundant information, the performance gap and modality under-optimization problem hinder effective multi-modal learning in practice, and are exacerbated in the confrontation of the missing data. To address this issue, we devise a novel **Cross-modal Prompting (ComP)** method, which emphasizes coherent information by enhancing modality-specific features and improves the overall recognition accuracy by boosting each modality’s performance. Specifically, a progressive prompt generation module with a dynamic gradient modulator is proposed to produce concise and consistent modality semantic cues. Meanwhile, cross-modal knowledge propagation selectively amplifies the consistent information in modality features with the delivered prompts to enhance the discrimination of the modality-specific output. Additionally, a coordinator is employed to dynamically re-weight the modality outputs as a complement to the balance strategy to improve the model’s efficacy. Extensive experiments on 4 datasets with 7 SOTA methods under different missing rates validate the effectiveness of our proposed method.

**Code** — <https://github.com/WenjueHE/2026-AAAI-ComP>

## Introduction

The growing demand for human-centric artificial intelligence has stimulated the development of emotion recognition (ER), which is considered as one of the most fundamental tasks in sentiment analysis. Compared to single-modal emotion recognition, multi-modal emotion recognition (MER) is able to deal with more complicated cases such as satires and metaphors, and has been widely applied to depression detection (Fan et al. 2024; Tao et al. 2024), dialogue systems (Firdaus et al. 2020; Liang et al. 2022), and so on.

However, the aforementioned traditional MER methods are generally based on an assumption that instances from all modalities are fully observed, which is often broken by limitations in data collection, transmission, and storage in real-world scenarios (Lian et al. 2023; He, Zhang, and Zhu 2025;

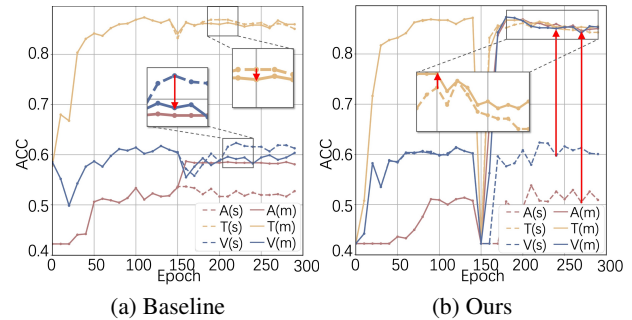


Figure 1: Modality-specific ACCs of single-modal (s) and multi-modal (m) training of 3 modalities. **(a)** In baseline methods, the performance of each modality varies, and the Video (V) and Text (T) modalities suffer a degradation after multi-modal co-training. **(b)** By cross-modal prompting, all modalities benefit from multi-modal learning.

Zhang and He 2023). For instance, background noise might disable the audio modality, and severe accents may lead to failures in speech recognition when constructing the text modality. The incomplete data problem impairs the cross-modal consistency, making it more challenging to achieve accurate emotion recognition.

Several attempts have been made to deal with the incomplete MER problem. Zhao *et al.* (2021) employ the cascaded residual autoencoders (CRAs) to perform conversion from the observed data to the missing ones. Taking the data distribution of each modality into consideration, Wang *et al.* (2023) propose to recover the missing instances with the diffusion model by a noising-denoising process. Nevertheless, recovering missing data is resource-consuming, and not all the recovered information is necessarily utilized in the training stage. Instead, another stream of methods seeks to learn a unified representation with potential missing modalities. Guo *et al.* (2024) propose different types of prompts to make up for the missing modality information. Li *et al.* (2024) perform multi-grained alignment between the teacher branch pre-trained on the complete data and the student branch trained on the incomplete data. Inspired by the idea of mixture of experts, Xu *et al.* (2024) feed the modality-specific features to all experts to leverage their knowledge and design

\*Corresponding Author: Zheng Zhang

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

a router to determine their weights.

Although progress has been made in IMER, modalities in vanilla multi-modal methods are not fully explored due to the modality imbalance problem (Wei et al. 2025), which could be summarized into modality performance gap, *i.e.*, performances of different modalities vary greatly, and modality under-optimization, *i.e.*, modality-specific performance degrades after multi-modal co-training. Fig. 1(a) presents a showcase of the modality imbalance problem. Such phenomena could be attributed to the intrinsic heterogeneity of multi-modal data. On one hand, the features of some modalities could be less focused on the task-relevant information than others. On the other hand, the features lie in different spaces, and directly concatenating them together may lead to conflicts. Existing IMER methods generally ignore such discrepancies and treat each modality independently before fusion, therefore fail to address the modality imbalance problem. Contrary to the existing works, in this paper, we propose to first enhance the representation of each modality by its cross-modal counterparts through knowledge propagation. In this way, cross-modal information is delivered to each modality branch to emphasize the emotion-related information that is consistent across modalities, thereby suppressing the misleading information and encouraging the non-dominant modalities to be better utilized. For more effective consistency enhancement, two questions need to be carefully considered:

**Q1: What kind of knowledge should be passed to other modalities?**

**Q2: How can each modality combine its own information with that of the others?**

For the first question, inspired by recent progress in prompt learning (Zhou et al. 2022b,a), we design a progressive prompt generation module with momentum updating to learn representative and consistent prompts. Specifically speaking, the global input is compressed into a small number of prototypes with gradient modulation to avoid the prototypes being dominated by the easy samples, which are fused with the contextual features to obtain low-dimensional representative prompts. For the second question, a simple yet effective cross-modal knowledge propagation module is designed to conduct mutual updating for the modality feature and prompts for invariant information enhancement. Meanwhile, this measure seamlessly reconstructs the incomplete samples without any extra processes. A coordinator is introduced subsequently to rebalance the modality outputs. Combining the components above, the general framework of our proposed method is illustrated in Fig. 2. As demonstrated in Fig. 1(b), our ComP method well alleviates the modality imbalance problem and boosts the performance of modalities before feeding them into the coordinator.

Our main contributions are listed as follows:

1. We propose a novel balanced learning scheme for incomplete multi-modal emotion recognition by promoting each modality to communicate with others. To the best of our knowledge, this is one of the first attempts to alleviate the imbalanced incomplete multi-modal learning problem by prompt learning.
2. To ensure the quality of broadcasted knowledge, we propose a novel prototype-based progressive prompt generation method with sample-level gradient adjustment, which guarantees the representativeness and consistency of the learned prototypes.
3. Extensive experiments validate that our proposed method well balances all modalities and reaches superior performance over the existing SOTA methods by promoting each specific modality.

## Related Works

### Incomplete Multi-modal Emotion Recognition

Compared to early attempts that are based on the assumption of full data accessibility, recent studies have delved into learning from incomplete modalities. Some methods try to directly recover the missing instance. Deng *et al.* (2025) model the correlations among features by graphs and imputes the missing instances with their k-nearest-neighbors of the rest observed modalities. Wu *et al.* (2024) maintain a feature queue for each modal to simulate the data distribution and conduct cross-modality recovery from both sample and distribution perspectives. However, directly restoring the original data could be less efficient, as the training process could be resource-consuming, and only a partial restoration of information could contribute to the recognition task. Another stream of methods explores the cross-modal connections and learn a consistent representation to mitigate the influence of incompleteness. Lian *et al.* (2023) design graph neural networks to simultaneously model the temporal correlation as well as the speaker-wise dependencies. Inspired by the idea of mixture of experts, MoMKE (Xu, Jiang, and Liang 2024) feeds the feature of each modality into several experts and designs a router to dynamically adjust their weights. However, in such methods, the performances of some modalities are generally inferior to others, which results in under-exploration of given data and suboptimal results.

### Modality-balanced Multi-modal Learning

In multi-modal learning, the model’s performance may rely greatly on a single dominant modality (Zou, Huang, and Shen 2023; Zhang, Wang, and Yu 2024). In some cases, adding an additional modality has little contribution to the overall performance (Vielzeuf et al. 2018), and may even impair it (Wang, Tran, and Feiszli 2020). To address these issues, balanced multi-modal learning methods seek to fully utilize the potential of each modality and reach higher performance. Structure-based balance learning methods design specific modules to guide the weaker modalities to learn towards the stronger ones. Yang *et al.* (2024) propose to distill the learnable multi-modal encoders from a pre-trained single modality encoder to learn representative and consistent features for vision-language retrieval. Fan *et al.* (2023) accelerate the learning of slow-learning modalities towards the class prototypes and introduce an entropy regularization to alleviate the suppression from the stronger modalities. Gradient-based methods, on the other hand, dynamically adjust the gradient for certain variables to ensure suf-

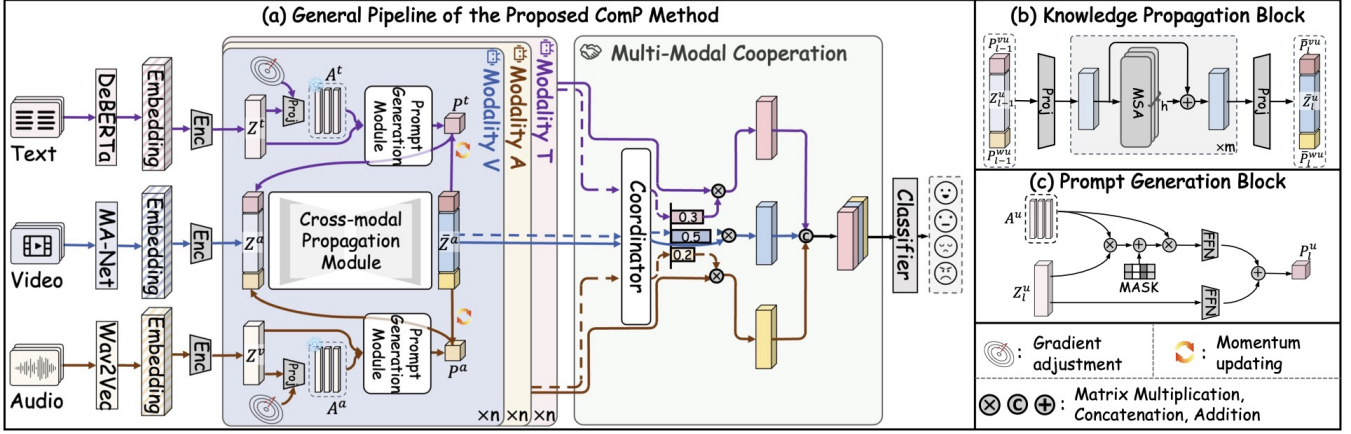


Figure 2: (a) The general framework of the proposed ComP method. First, the prompt generation (PG) module, which consists of several PG blocks (c), compresses the modality-specific features to obtain consistent yet representative prompts. Then, the prompts are passed to other modalities through the knowledge propagation (KP) module (b) to enhance the task-relevant information. A multi-modal cooperation module (Cr) further examines the significance of each modality and reweighs them before fusion.

efficient training for the weaker modalities. BML (Wei et al. 2025) suppresses the dominant modality by randomly dropping some of its features as well as decreasing its learning rate to encourage other modalities' training. Wei *et al.* (2024) borrow the idea of Pareto efficiency to measure the effect of each modality and avoid gradient conflicts by adjusting the gradient direction.

The abovementioned methods generally focus on the complete modality scenario, and ignore the information loss caused by the unobserved instances that could further exacerbate the modality imbalance problem. On the other hand, our proposed method integrates missing instance management and cross-modality balancing into a unified framework of consensus information enhancing by progressive prompt generation and knowledge propagation, efficiently alleviates the performance gap and modality underoptimization problem in IMER.

## The Proposed Method

### General Framework

Following existing works (Lian et al. 2023; Xu, Jiang, and Liang 2024), for each utterance, the audio, text, and video embeddings, denoted as  $\mathbf{X}^a$ ,  $\mathbf{X}^t$ , and  $\mathbf{X}^v$ , are first extracted with Wav2Vec (Baeveski et al. 2020), DeBERTa (He et al. 2021), and MA-Net (Zhao, Liu, and Wang 2021), respectively. For clarity, we use  $\mathcal{M} = \{a, t, v\}$  to represent all modalities that occur in the experiments, and superscripts  $u, v, w \in \mathcal{M}$  to represent certain modalities, which satisfy  $u \neq v \neq w$ .

For the incomplete problem, instances in some modalities could be unobserved, which could be suggested by the following indicator:

$$\gamma_i^u = \begin{cases} 1, & \mathbf{x}_i^u \in \mathcal{X}_o, \\ 0, & \mathbf{x}_i^u \in \mathcal{X}_m. \end{cases} \quad (1)$$

In eq. (1),  $\mathcal{X}_o$  and  $\mathcal{X}_m$  denote the set consisting of the observed and missing instances, respectively.  $\hat{\mathbf{X}} = \{(\gamma^u \mathbf{1}_d^\top) \odot \mathbf{X}^u | u \in \mathcal{M}\}$  imputes the missing instances with 0 for the subsequent processes to eliminate the noise caused by incompleteness, where  $\mathbf{1}_d^\top \in \mathbb{R}^{1 \times d}$  is a vector with all elements equals to 1,  $\odot$  represents Hadamard product.

Our goal is to build a function  $C : \hat{\mathcal{X}} \rightarrow \mathbf{Y}'$  that maps the incomplete feature  $\hat{\mathcal{X}}$  to the emotion predictions  $\mathbf{Y}'$ . To achieve this, the model follows a two-stage training scheme. In the first stage, a feature updating function  $f^u$  and an encoder  $\text{Enc}^u$  are jointly trained on each specific modality by minimizing the following objective:

$$\mathcal{L} = \sum_{u \in \mathcal{M}} \mathcal{L}_{\text{enc}}(\mathbf{Z}^u, \hat{\mathbf{X}}^u) + \sum_{u \in \mathcal{M}} \mathcal{L}_{\text{task}}(\mathbf{Y}^u, \mathbf{Y}), \quad (2)$$

where  $\mathcal{L}_{\text{enc}}$  and  $\mathcal{L}_{\text{task}}$  is the reconstruction and task-relevant loss, respectively,  $\mathbf{Z}^u = \text{Enc}^u(\hat{\mathbf{X}}^u) \in \mathbb{R}^{n \times d}$  is features extracted by encoders, and  $\mathbf{Y}^u = \text{Classifier}^u(f^u(\mathbf{Z}^u)) \in \mathbb{R}^{n \times 1}$  is the modality-specific predictions.

However, due to the intrinsic characteristics of multi-modal data, the capability of each function  $f^u$  to recognize emotions is different. To alleviate such an imbalance and make full use of the multi-modal data, in the second stage, the feature of each modality is compressed to a prompt by the *prompt generation module*, and passed into the *cross-modal knowledge propagation module* of other modalities to enhance the task-relevant consensus information and boost modality performance. Finally, the output of each modality is weighted by a *multi-modal cooperation module* before fusing them together. A detailed explanation of each module is given in the rest of this section.

### Cross-modal Knowledge Propagation

To address the modality-imbalance problem that the learning capability of some modalities is inferior to others, we enhance each modality by exposing it to the compressed data

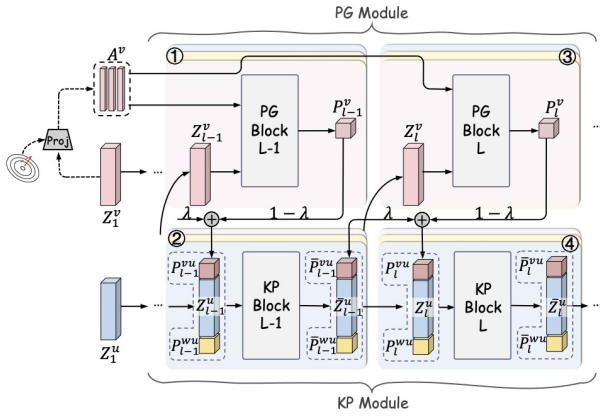


Figure 3: Interactions between the prompt generation (PG) blocks and knowledge propagation (KP) blocks.

from the other modalities. As the emotion signals are consistent across modalities, such a measure could effectively stress the task-relevant information and boost the cross-modal synergistic effect.

Specifically, the knowledge propagation module consists of  $n$  corresponding blocks (Fig. 2(b)). At the beginning of the  $l$ -th block in modality  $u$ , semantical prompts  $P_l^{vu}, P_l^{wu} \in \mathbb{R}^{n \times p}$  from modalities  $v$  and  $w$  (which will be introduced in the next subsection) are passed and concatenated with the block input  $Z_l^u$ , where we define  $Z_1^u = Z^u$  for clarity. The concatenated features are first projected to a lower-dimensional fused feature  $G_{l,0}^u$ :

$$G_{l,0}^u = \text{Linear}_{(d+2p) \rightarrow d}(\text{Concat}([Z_l^u, P_l^{vu}, P_l^{wu}])). \quad (3)$$

The compressed features are then passed into  $m$  multi-head self-attention (MSA) layers with residual connection to further enhance the cross-modal communication. Missing instances are masked in MSA to avoid extra noise:

$$G_{l,k}^u = G_{l,k-1}^u + \text{MSA}(G_{l,k-1}^u), \quad 1 \leq k \leq m, \quad (4)$$

where  $G_{l,k}^u \in \mathbb{R}^{n \times d}$  denotes the output of the  $k$ -th layer in the  $l$ -th block. Finally, the compressed features are projected back to the original space to obtain the updated features and prompts:

$$[\bar{Z}_l^u, \bar{P}_l^{vu}, \bar{P}_l^{wu}] = \text{Linear}_{d \rightarrow (d+2p)}(G_{l,m}^u). \quad (5)$$

$\bar{Z}_l^u \in \mathbb{R}^{n \times d}$ ,  $\bar{P}_l^{vu} \in \mathbb{R}^{n \times p}$ , and  $\bar{P}_l^{wu} \in \mathbb{R}^{n \times p}$  represents the renewed view-specific feature and prompts passed to modality  $u$ , respectively. For the next block,  $\bar{Z}_l^u$  is directly treated as the modality feature, *i.e.*,  $Z_{l+1}^u = \bar{Z}_l^u$ , while  $\bar{P}_l^{vu}$  and  $\bar{P}_l^{wu}$  are further processed before usage. Additionally, after this process, the missing instances are also seamlessly reconstructed by their cross-modal counterparts.

### Prompt Generation

Prompts carry the compressed modality information, and play a crucial role in the cross-modality knowledge propagation. To effectively highlight the consensus information, the intrinsic emotional information in each modality should

be reflected by the prompts, which requires the prompts to be dynamically updated with the features. Additionally, the inter-block consistency should be preserved to avoid extra noise caused by abrupt changes in communication. In this section, we present a novel progressive prompt generation method to construct informative and consistent prompts.

**Original Information Preservation.** Learning a small set of prototypes to represent the distribution of the original data has been proven effective and resource-efficient (Li et al. 2021; Wei et al. 2023). Traditional prototype learning methods generally employ a fixed generation scheme such as  $k$ -means or averaging, which could be less representative as it cannot be updated. Contrary to the existing methods, we conduct compression on the batch size dimension and reduce the number of instances to the number of prototypes:

$$A^u = (\text{MLP}(Z^{u\top}))^\top, \quad (6)$$

where  $A^u \in \mathbb{R}^{c \times d}$  is the learned prototype matrix with  $c$  instances, and each modality-specific Multilayer Perceptron (MLP) is consisted of a  $n$ -by- $c$  linear layer  $g^u(Z^{u\top}) = Z^{u\top} M^u + b^u$ , a GeLU activation layer, a  $c$ -by- $c$  linear layer, and a dropout layer. The prototype  $A^u$  represents the characteristics of the original feature  $Z^u$ , and is frozen among all knowledge propagation blocks within each batch.

**Prompt Update.** As illustrated in Fig. 2(c), the obtained prototype  $A^u$  is used to represent the global input features  $Z^u$ . For the  $l$ -th prompt generation block, given the output of the previous layer  $\bar{Z}_{l-1}^u$  and  $\bar{P}_{l-1}^u$ , the prompts are generated by:

$$S_l^u = \text{softmax}(\text{Sim}(Z_l^u, A^u) + \text{Mask}), \quad (7)$$

$$\tilde{P}_l^u = \text{MLP}_p(\text{FFN}_1(S_l^u A^u) + \text{FFN}_2(\bar{Z}_{l-1}^u)), \quad (8)$$

where  $\text{Sim}$  denotes the cosine similarity,  $\text{Mask} \in \mathbb{R}^{n \times c}$  prevents the missing instances in  $Z_l^u$  from disturbing the results by setting the missing instance related entries to a large negative value,  $\text{FFN}_1$  and  $\text{FFN}_2$  are two feed-forward networks (FFNs),  $\text{MLP}_p$  projects the feature to  $p$ -dim. Considering that after the communication process, the prompts passed to other views have also been updated, the final prompts are obtained in a momentum updating manner:

$$P_l^{vu} = \lambda \bar{P}_{l-1}^{vu} + (1 - \lambda) \tilde{P}_l^u, \quad (9)$$

where  $0 \leq \lambda \leq 1$  is a hyper-parameter. In this way, the prompts are updated more stably.

**Instance-level Gradient Adjustment.** Information in some samples could be more conspicuous than the others (Yang et al. 2025; Wei et al. 2024). To avoid the prototypes being dominated by the easy samples, we propose to dynamically quantify the difficulty of each sample by its dependency on cross-modal assistance, and adjust their learning rates to equilibrate the contribution of both simple and challenging samples.

We first calculate the logit error of modality  $u$ , *i.e.*,  $E^u$ , which quantifies the distance between the ground truth and the prediction with the index  $y$  of the ground truth (gt):

$$\begin{aligned} \text{logit}^u &= \text{Proj}^u(\bar{Z}_l^u)_y, \\ E^u &= |\text{gt} - \text{logit}^u|. \end{aligned}$$

Missing Rate		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Dataset	Method	ACC%/UA(%)	ACC%/UA(%)	ACC%/UA(%)	ACC%/UA(%)	ACC%/UA(%)	ACC%/UA(%)	ACC%/UA(%)
IEMOCAPFour	MMIN(ACL '21)	72.30/72.85	69.22/70.28	66.21/66.80	62.77/62.99	59.89/59.55	57.87/57.65	54.23/53.49
	GCNet(TPAMI '23)	74.82/74.21	75.87/75.59	74.49/74.31	74.43/73.96	72.67/72.34	72.65/73.33	71.00/70.26
	MoMKE(MM '24)	76.70/75.80	75.25/74.02	73.47/72.59	71.42/70.25	69.73/68.95	67.95/66.60	66.52/65.53
	SDR-GNN(KBS '25)	78.48/78.25	77.83/78.63	78.22/78.29	76.65/76.32	75.47/75.85	73.87/74.45	70.52/72.21
	Ours	<b>80.66/81.09</b>	<b>79.58/80.22</b>	<b>78.37/79.16</b>	<b>77.21/77.96</b>	<b>75.62/76.44</b>	<b>74.28/75.27</b>	<b>73.41/74.09</b>
	$\Delta$ SOTA	$\uparrow$ 2.18/2.84	$\uparrow$ 1.77/1.39	$\uparrow$ 0.15/0.87	$\uparrow$ 0.56/1.64	$\uparrow$ 0.15/0.59	$\uparrow$ 0.43/0.08	$\uparrow$ 2.41/1.88
IEMOCAPSix	MMIN(ACL '21)	55.21/53.70	52.00/50.54	50.25/47.99	47.51/44.70	43.79/40.91	41.41/37.98	39.47/35.02
	GCNet(TPAMI '23)	57.44/57.94	56.81/54.84	55.38/53.33	56.44/54.86	<u>56.12/54.12</u>	<u>54.14/53.62</u>	<u>52.59/52.16</u>
	MoMKE(MM '24)	<u>60.54/58.36</u>	58.47/56.67	55.97/53.37	53.85/52.02	51.93/49.03	49.46/47.72	48.89/46.64
	SDR-GNN(KBS '25)	60.26/59.26	58.87/58.78	58.80/58.13	<b>57.64/56.71</b>	53.96/53.07	53.18/52.01	51.42/50.29
	Ours	<b>62.02/61.41</b>	<b>60.53/58.79</b>	<b>59.67/58.33</b>	<b>57.66/56.22</b>	<b>56.50/54.76</b>	<b>55.23/53.86</b>	<b>54.58/53.01</b>
	$\Delta$ SOTA	$\uparrow$ 1.48/2.15	$\uparrow$ 1.66/0.01	$\uparrow$ 0.87/0.20	$\uparrow$ 0.02/0.49	$\uparrow$ 0.38/0.64	$\uparrow$ 1.09/0.24	$\uparrow$ 1.99/0.85

Table 1: Experimental results on IEMOCAPFour and IEMOCAPSix. Best and second best results **boldfaced** and underlined.

The required modulation of a modality’s learning speed is then adjusted based on its relative learning difficulty:

$$\mathbf{W}_i^u = \frac{1}{2} \sum_{v \neq u} \frac{\sum_{j \neq i} (\mathbf{E}_j^u - \mathbf{E}_j^v)}{\sum_j (\mathbf{E}_j^u - \mathbf{E}_j^v)}. \quad (10)$$

$\mathbf{E}^u - \mathbf{E}^v$  is the relative error between modality  $u$  and  $v$ .  $\mathbf{W}^x \in \mathbb{R}^{n \times 1}$  is multiplied to the gradient of  $\mathbf{M}^x$  to adjust its learning rate:

$$\mathbf{M}_{t+1}^u = \mathbf{M}_t^u - \eta \hat{\mathbf{W}}^u \odot \nabla_{\mathbf{M}^u} \mathcal{L}(\mathbf{M}_t^u), \quad (11)$$

where  $\hat{\mathbf{W}}^x = \mathbf{W}^x \mathbf{1}_d^\top \in \mathbb{R}^{n \times d}$  is the weighting matrix,  $\eta$  is the original learning rate,  $\odot$  denotes the Hadamard product. In this way, if the relative error of one modality is large, indicating a greater need for other modalities’ assistance, its learning rate is decreased correspondingly to give it more chances to pass detailed information to the prototype. Fig. 3 shows the details of the interactions between PG and KP blocks.

### Multi-modal Cooperation

After the previously mentioned uni-modal training and cross-modal propagation, the potential of modalities is further explored, and each modality achieves a relatively high performance. In this stage, the modalities cooperate with each other to learn a shared representation so that the model can cope with various missing cases. Unlike traditional fixed fusing methods such as concatenating or averaging that treat each modality equally, inspired by recent advances in multi-modal fusion (Gao et al. 2024; Xu, Jiang, and Liang 2024), we employ an extra coordinator to dynamically decide how much each modality should contribute to the common representation in different cases:

$$\omega = [\omega^a, \omega^t, \omega^v] = \text{MLP}(\text{Concat}([\bar{\mathbf{Z}}^a, \bar{\mathbf{Z}}^t, \bar{\mathbf{Z}}^v])), \quad (12)$$

$$\bar{\omega}^u = \text{softmax}(\omega^u) = \frac{\exp(\omega^u)}{\sum_{v \in \mathcal{M}} \exp(\omega^v)}, \quad (13)$$

where  $\bar{\mathbf{Z}}^u$  is the output of the KP module.

After obtaining the weighting parameter  $\omega_u \in \mathbb{R}^{n \times 1}$ , the knowledge from each modality is scaled and concatenated, and fed into a classifier to obtain the final prediction:

$$\mathbf{F}_i = \text{Concat}([\bar{\omega}_i^a \bar{\mathbf{z}}_i^a, \bar{\omega}_i^t \bar{\mathbf{z}}_i^t, \bar{\omega}_i^v \bar{\mathbf{z}}_i^v]), \quad (14)$$

$$\mathbf{Y}' = \text{Classifier}(\mathbf{F}). \quad (15)$$

For emotion classification tasks, the model is optimized with the cross-entropy (CE) loss:

$$\mathcal{L}_{\text{task}} = \text{CE}(\mathbf{Y}, \mathbf{Y}') = - \sum_{i=1}^n \mathbf{Y}_i \log \mathbf{Y}'_i. \quad (16)$$

Otherwise, for emotion recognition tasks, the mean square error (MSE) loss is employed:

$$\mathcal{L}_{\text{task}} = \text{MSE}(\mathbf{Y}, \mathbf{Y}') = \frac{1}{n} \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{Y}'_i)^2. \quad (17)$$

## Experiments

### Datasets

To comprehensively evaluate the performance of our proposed method, we conduct experiments on four publicly available emotion recognition datasets, *i.e.*, CMU-MOSI (Zadeh et al. 2016), CMU-MOSEI (Zadeh et al. 2018), IEMOCAPFour (Busso et al. 2008), and IEMOCAPSix. As all the datasets are originally complete, we follow existing works (Zhao, Li, and Jin 2021; Wang, Li, and Cui 2023) and test them under a synthesized missing condition by randomly removing a proportion of data. The missing rate is defined by  $\text{MR} = (\sum_{i=1}^m \mathcal{N}^i) / (mN)$ , where  $m$ ,  $N$ , and  $\mathcal{N}^i$  is the number of modalities, total samples, and missing samples in modality  $i$ . Consistent with (Lian et al. 2023), MR is ranged from 0.1 to 0.7 to ensure that each sample is observed in at least one modality.

### Experiments on Incomplete Data

The performance of our proposed ComP method, along with 7 state-of-the-art (SOTA) methods including MCTN (Pham et al. 2019), MMIN (Zhao, Li, and Jin 2021), GCNet (Lian et al. 2023), DiCMoR (Wang, Cui, and Li 2023), IMDer (Wang, Li, and Cui 2023), MoMKE (Xu, Jiang, and Liang 2024), and SDR-GNN (Fu et al. 2025), is tested in this section. For a fair comparison, all methods are performed on the same features extracted by pre-trained models as ComP. For the IEMOCAP dataset, weighted accuracy (ACC) and unweighted accuracy (UA) are adopted to evaluate the performance, while for CMU-MOSI and CMU-MOSEI, ACC and weighted average F1-score (F1) are employed. The test results are shown in Table 1 and Table 2.

Missing Rate		0.1	0.2	0.3	0.4	0.5	0.6	0.7
Dataset	Method	ACC(%) / F1(%)	ACC(%) / F1(%)	ACC(%) / F1(%)	ACC(%) / F1(%)	ACC(%) / F1(%)	ACC(%) / F1(%)	ACC(%) / F1(%)
CMU-MOSI	MCTN <sup>◇</sup> (AAAI '19)	78.50/78.40	75.70/75.60	71.20/71.30	67.60/68.00	64.80/65.40	62.50/63.80	59.00/61.20
	MMIN(ACL '21)	79.88/79.94	77.74/77.82	73.78/73.82	70.43/70.20	67.07/65.63	60.98/60.00	61.43/61.56
	GCNet(TPAMI '23)	80.95/81.02	78.51/78.53	77.59/77.39	73.32/73.48	74.39/74.46	65.70/65.91	64.79/64.99
	DiCMoR <sup>†</sup> (ICCV '23)	83.90/83.90	82.00/82.10	80.20/80.40	77.70/77.90	76.40/76.70	73.00/73.30	70.80/71.10
	IMDer <sup>‡</sup> (NIPS '23)	84.90/84.80	83.50/83.40	<u>81.20/81.00</u>	78.60/78.50	76.20/75.90	<u>74.70/74.00</u>	<u>71.90/71.20</u>
	MoMKE(MM '24)	<u>86.74/86.74</u>	<u>83.38/83.44</u>	80.64/80.72	77.90/78.02	76.68/76.80	73.93/74.08	70.58/70.76
	SDR-GNN(KBS '25)	82.47/82.32	81.40/81.19	80.80/81.40	<u>79.62/79.57</u>	<u>78.35/78.26</u>	69.98/69.82	69.32/69.21
	Ours	<b>87.20/87.11</b>	<b>85.37/85.23</b>	<b>83.69/83.67</b>	<b>81.10/81.18</b>	<b>79.27/78.96</b>	<b>75.61/75.74</b>	<b>73.17/73.33</b>
	ΔSOTA	↑ 0.46/0.37	↑ 1.87/1.79	↑ 2.49/2.27	↑ 1.48/1.61	↑ 0.92/0.70	↑ 0.91/1.66	↑ 1.27/2.13
	CMU-MOSEI	MCTN <sup>◇</sup> (AAAI '19)	81.60/81.80	78.70/79.00	76.20/76.90	74.10/74.30	72.60/73.60	71.10/73.20
MMIN(ACL '21)		83.82/83.63	82.20/81.90	79.94/79.27	78.59/77.69	75.92/75.73	73.61/71.98	74.35/72.73
GCNet(TPAMI '23)		85.80/85.82	<u>85.14/85.10</u>	<u>84.48/84.35</u>	<u>83.05/82.90</u>	82.03/82.08	<u>81.29/81.03</u>	<u>80.05/80.03</u>
DiCMoR <sup>†</sup> (ICCV '23)		83.50/83.70	81.50/81.80	79.30/79.80	77.40/78.70	75.80/77.70	73.70/76.70	72.20/75.40
IMDer <sup>‡</sup> (NIPS '23)		84.80/84.60	82.70/82.40	81.30/80.70	79.30/78.10	79.00/77.40	78.00/75.50	77.30/74.60
MoMKE(MM '24)		<u>85.88/85.65</u>	84.76/84.60	82.61/82.42	80.99/80.77	79.09/78.93	77.49/76.92	76.00/75.86
SDR-GNN(KBS '25)		85.75/85.63	85.01/85.06	<u>84.25/84.37</u>	82.17/82.17	<u>82.12/82.25</u>	80.59/81.18	79.97/80.54
Ours		<b>86.60/86.48</b>	<b>85.83/85.59</b>	<b>85.28/85.10</b>	<b>84.20/83.94</b>	<b>83.30/83.12</b>	<b>82.17/81.78</b>	<b>80.82/80.55</b>
ΔSOTA		↑ 0.72/0.66	↑ 0.69/0.49	↑ 0.80/0.73	↑ 1.15/1.04	↑ 1.18/0.87	↑ 0.88/0.60	↑ 0.77/0.01

Table 2: Experimental results on CMU-MOSI and CMU-MOSEI. Best and second best results are **boldfaced** and underlined. Results with <sup>†</sup>, <sup>‡</sup> and <sup>◇</sup> are from (Wang, Cui, and Li 2023), (Wang, Li, and Cui 2023), and (Lian et al. 2023).

The recognition accuracies of all methods decline as the MR increases due to the uncertainty and noise introduced by incompleteness. DiCMoR and IMDer restore the original shallow features with generative models, so that traditional MER methods built on complete datasets could be transferred. However, such reconstruction could be noisy and less representative, resulting in relatively inferior performance even at a low MR, with a rapid degradation as MR increases. GCNet employs an early-fusion strategy by concatenating modality features and passing them through an LSTM, which fails to comprehensively investigate the cross-modality correlation at the utterance level and leads to less accurate recognition. MoMKE performs cross-modal communication at the model level by passing features to experts pre-trained independently on single modalities. However, such coarse-grained message passing fails to comprehensively exchange feature-level messaging, and its testing results drop severely when the MR becomes large.

Compared to the existing methods, our knowledge propagation module with prompt learning enables more thorough communication across views by mutual information enhancement, which alleviates the side-effect caused by missing instances and improves the overall performance through single-modal boosting. Our ComP almost always outperforms other methods on 4 datasets with different missing rates. Additionally, the increasing number of missing instances has a relatively small impact on our methods than on others, which again validates that our coherency enhancement strategy could effectively extract the task-relevant information even when some instances are missing.

## Ablation Study

Table 3 demonstrates the importance of each component. When all components are removed, the output of the encoders is directly concatenated for classification, while when all components are added, the model is equivalent to ComP. **1) Knowledge propagation (KP)** is the basis of our proposed method. When KP is removed, the subsequent com-

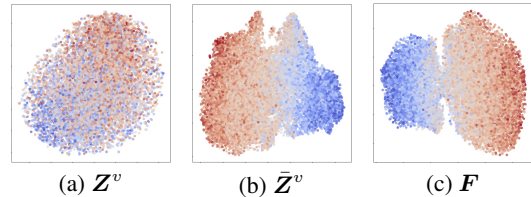


Figure 4: T-SNE visualization results of  $Z^v$ ,  $\bar{Z}^v$ , and  $F$  on CMU-MOSEI dataset. Positive and negative emotions are marked by red and blue, respectively.

ponents are disabled, and the general performance therefore suffers a severe drop, especially when the MR is high. **2) Prompt generation (PG)** aims at generating informative and representative prompts. Removing it adds noise to the propagated knowledge, and also results in a relatively high degradation. **3) By adding modality coordination (Cr)**, the performance of the model continues to increase, indicating that although the modality imbalance problem is alleviated, the employed coordinator could dynamically adjust the importance of each modality. **4) Gradient modulation (GM)** further improves the prototype learning in prompt generation, and its promotion effect is more pronounced on CMU-MOSI. The reason could be its various emotion intensity as samples with stronger emotions could easily dominate prototype learning in CMU-MOSI without GM.

## Visualization

To further examine the effectiveness of our proposed knowledge propagation strategy, T-SNE visualization of  $Z^v$ ,  $\bar{Z}^v$ , and  $F$ , which represent the primary features before propagation, the modality-specific features after propagation, and the fused features after coordination, are illustrated in Fig. 4. Different colors represent the emotion spectrum, where saturated red and blue represent the most positive (score = 3) and negative (score = -3), respectively. The features before propagation, illustrated in Fig. 4(a), are non-separable. Af-

Dataset	Tested Components				Missing Rate													
					0.1		0.2		0.3		0.4		0.5		0.6		0.7	
	KP	PG	Cr	GM	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)	Acc(%)	F1(%)
CMU-MOSI	○	○	○	○	83.69	83.79	81.86	81.97	78.96	79.09	75.30	75.44	72.10	72.27	69.97	70.07	68.75	68.90
	●	○	○	○	85.67	85.63	83.23	83.13	81.40	81.27	78.51	78.62	75.15	75.29	73.17	73.29	70.88	70.96
	●	●	○	○	85.82	85.76	83.69	83.73	82.01	81.98	78.96	79.05	75.46	75.61	73.48	73.55	71.65	71.81
	●	●	●	○	86.13	86.09	84.15	84.20	82.32	82.42	79.57	79.70	76.52	76.65	75.00	75.09	71.80	71.97
	○	●	●	●	84.15	84.16	81.86	81.91	79.73	79.59	77.59	77.69	73.78	73.80	71.04	71.20	69.97	69.99
	●	○	●	●	85.82	85.73	83.69	83.66	82.32	82.28	79.42	79.48	76.52	76.66	73.32	73.44	71.19	71.09
	●	●	○	●	86.13	86.14	83.99	83.96	82.93	82.88	79.42	79.46	77.29	77.32	74.09	74.24	72.10	72.14
	●	●	●	●	<b>87.20</b>	<b>87.11</b>	<b>85.37</b>	<b>85.23</b>	<b>83.69</b>	<b>83.67</b>	<b>81.10</b>	<b>81.18</b>	<b>79.27</b>	<b>78.96</b>	<b>75.61</b>	<b>75.74</b>	<b>73.17</b>	<b>73.33</b>
IEMOCAPFour	○	○	○	○	72.12	71.82	70.51	69.89	68.36	67.86	65.89	65.35	62.71	62.38	59.97	59.82	57.85	57.58
	●	○	○	○	78.94	79.74	78.42	79.02	76.85	77.50	75.77	76.39	73.71	74.71	72.44	72.93	71.83	72.72
	●	●	○	○	79.70	80.64	79.04	79.73	77.51	78.43	76.56	77.35	75.20	75.89	73.47	74.53	72.99	73.74
	●	●	●	○	80.57	81.03	79.24	80.17	78.06	78.52	76.74	77.51	75.35	76.40	73.73	74.80	73.03	74.03
	○	●	●	●	72.09	72.16	70.60	70.18	68.13	67.24	65.49	64.76	63.08	62.32	59.57	59.59	57.36	57.05
	●	○	●	●	79.20	79.63	78.71	79.72	77.63	78.70	75.55	76.22	73.65	74.57	72.30	73.51	71.86	73.32
	●	●	○	●	79.70	80.64	78.75	79.44	78.11	78.66	76.61	77.82	75.00	75.29	73.86	74.27	72.38	72.70
	●	●	●	●	<b>80.66</b>	<b>81.09</b>	<b>79.58</b>	<b>80.22</b>	<b>78.37</b>	<b>79.16</b>	<b>77.21</b>	<b>77.96</b>	<b>75.62</b>	<b>76.44</b>	<b>74.28</b>	<b>75.27</b>	<b>73.41</b>	<b>74.09</b>

Table 3: Ablation study on 2 datasets. ○ and ● denote removed and preserved components, respectively. Best results **boldfaced**.

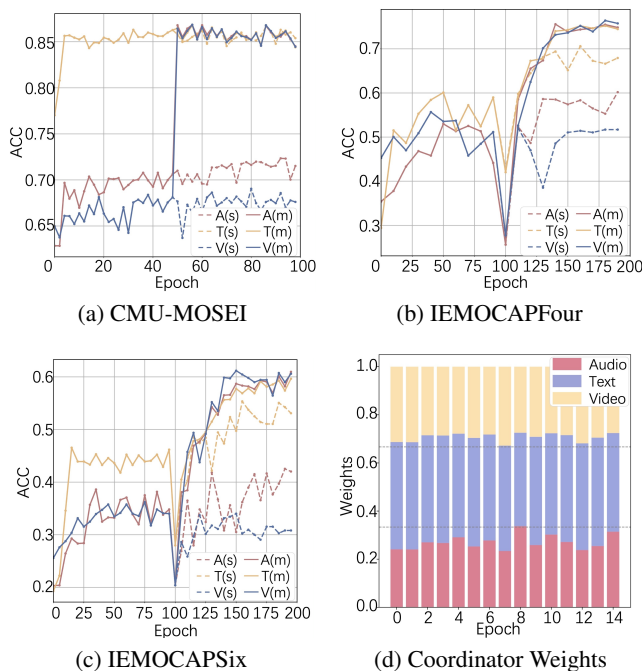


Figure 5: (a-c) Modality-specific accuracies versus training epochs on 3 tested datasets. (d)  $\bar{\omega}_a$ ,  $\bar{\omega}_t$ , and  $\bar{\omega}_v$  for different utterances on CMU-MOSI.

ter knowledge propagation, the sentiment-related common information is enhanced across all modalities, which can be evidenced by the structured organization of the dots in Fig. 4(b). Additionally, the overlap between the two parts is further narrowed in Fig. 4(c), demonstrating an enhanced separability of the features achieved by modality cooperation.

## Modality-balance Study

**Modality-specific Accuracy.** The modality-specific performances of our ComP method on three datasets are illustrated in Fig. 5(a)-(c). The single modality performances

represented in dotted lines vary across all the datasets. Contrarily, the solid lines indicating the single modality performance during co-training experience a sharp improvement at the beginning of the second training stage, where the cross-modal knowledge propagation is involved in the training process. Additionally, for the same modality, the solid line consistently lies above its dotted counterpart, which indicates that our training strategy not only fully explores the potential of each modality, but also jointly utilizes the complementary information from different modalities to reach better emotion recognition performance.

**Coordinator Weights.** The weights generated by the coordinator of the utterances from the same conversation are illustrated in Fig. 5(d). On one hand, the weights exhibit a relatively balanced distribution across views, validating that the discriminative ability of features is comparable across views after knowledge propagation. On the other hand, the weights of the text modality are slightly higher than the other modalities, which showcases the preserved intrinsic of the multi-modal data as well as the necessity of the coordinator.

## Conclusion

In this paper, we presented a novel Cross-modal Prompting (ComP) method to address the modality imbalance problems including modality performance gap and modality under-optimization in IMER by task-relevant consensus information enhancement through cross-modality knowledge propagation. Specifically speaking, a global prototype based prompt generation module with gradient regulation was proposed to compress the modality-specific knowledge in a consistent way. Afterward, the consensus information in each modality was enhanced by interactions with cross-modality prompts, so that the accuracy of single modality recognition significantly improves, with the aforementioned modality imbalance problems alleviated. Additionally, a coordinator was designed to adaptively adjust the importance of modality fusion, which further ensures the discriminative capability of the proposed method. Experiments validated that our ComP method not only improves the overall performance, but further explores the potential of each modality.

## Acknowledgments

This research is partially supported by National Natural Science Foundation of China (Grant no. 62372132) and Shenzhen Science and Technology Program (Grant no. RCYX20221008092852077). The authors would also like to thank Huawei Ascend Cloud Ecological Development Project for providing high-performance Ascend 910 processors.

## References

- Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 12449–12460.
- Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42: 335–359.
- Deng, Y.; Bian, J.; Wu, S.; Lai, J.; and Xie, X. 2025. Multiplex graph aggregation and feature refinement for unsupervised incomplete multimodal emotion recognition. *Information Fusion*, 114: 102711.
- Fan, H.; Zhang, X.; Xu, Y.; Fang, J.; Zhang, S.; Zhao, X.; and Yu, J. 2024. Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. *Information Fusion*, 104: 102161.
- Fan, Y.; Xu, W.; Wang, H.; Wang, J.; and Guo, S. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20029–20038.
- Firdaus, M.; Chauhan, H.; Ekbal, A.; and Bhattacharyya, P. 2020. EmoSen: Generating sentiment and emotion controlled responses in a multimodal dialogue system. *IEEE Transactions on Affective Computing*, 13(3): 1555–1566.
- Fu, F.; Ai, W.; Yang, F.; Shou, Y.; Meng, T.; and Li, K. 2025. SDR-GNN: Spectral domain reconstruction graph neural network for incomplete multimodal learning in conversational emotion recognition. *Knowledge-Based Systems*, 309: 112825.
- Gao, Z.; Jiang, X.; Xu, X.; Shen, F.; Li, Y.; and Shen, H. T. 2024. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26876–26885.
- Guo, Z.; Jin, T.; and Zhao, Z. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 1726–1736.
- He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced bert with disentangled attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 1–14.
- He, W.-J.; Zhang, Z.; and Zhu, X. 2025. Dual-Correlation-Guided Anchor Learning for Scalable Incomplete Multi-View Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 17336 – 17349.
- Li, G.; Jampani, V.; Sevilla-Lara, L.; Sun, D.; Kim, J.; and Kim, J. 2021. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8334–8343.
- Li, M.; Yang, D.; Liu, Y.; Wang, S.; Chen, J.; Wang, S.; Wei, J.; Jiang, Y.; Xu, Q.; Hou, X.; et al. 2024. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1–15.
- Lian, Z.; Chen, L.; Sun, L.; Liu, B.; and Tao, J. 2023. G-net: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8419–8432.
- Liang, Y.; Meng, F.; Zhang, Y.; Chen, Y.; Xu, J.; and Zhou, J. 2022. Emotional conversation generation with heterogeneous graph neural network. *Artificial Intelligence*, 308: 103714.
- Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, 6892–6899.
- Tao, Y.; Yang, M.; Li, H.; Wu, Y.; and Hu, B. 2024. DepM-STAT: Multimodal spatio-temporal attentional transformer for depression detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(7): 2956–2966.
- Vielzeuf, V.; Lechervy, A.; Pateux, S.; and Jurie, F. 2018. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 1–15.
- Wang, W.; Tran, D.; and Feiszli, M. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12695–12705.
- Wang, Y.; Cui, Z.; and Li, Y. 2023. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22025–22034.
- Wang, Y.; Li, Y.; and Cui, Z. 2023. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1–12.
- Wei, Y.; Feng, R.; Wang, Z.; and Hu, D. 2024. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27338–27347.
- Wei, Y.; and Hu, D. 2024. MMPareto: boosting multimodal learning with innocent unimodal assistance. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1–11.

- Wei, Y.; Hu, D.; Du, H.; and Wen, J.-R. 2025. On-the-fly modulation for balanced multimodal learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1): 469–485.
- Wei, Y.; Ye, J.; Huang, Z.; Zhang, J.; and Shan, H. 2023. Online prototype learning for online continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 18764–18774.
- Wu, D.; Yang, D.; Zhou, Y.; and Ma, C. 2024. Robust multimodal sentiment analysis of image-text pairs by distribution-based feature recovery and fusion. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 5780–5789.
- Xu, W.; Jiang, H.; and Liang, X. 2024. Leveraging Knowledge of Modality Experts for Incomplete Multimodal Learning. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 438–446.
- Yang, Y.; Pan, H.; Jiang, Q.-Y.; Xu, Y.; and Tang, J. 2025. Learning to rebalance multi-modal optimization by adaptively masking subnetworks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4553–4566.
- Yang, Y.; Xi, W.; Zhou, L.; and Tang, J. 2024. Rebalanced vision-language retrieval considering structure-aware distillation. *IEEE Transactions on Image Processing*, 6881–6892.
- Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88.
- Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. 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 (ACL)*, 2236–2246.
- Zhang, H.; Wang, W.; and Yu, T. 2024. Towards robust multimodal sentiment analysis with incomplete data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1–12.
- Zhang, Z.; and He, W.-J. 2023. Tensorized topological graph learning for generalized incomplete multi-view clustering. *Information Fusion*, 100: 101914.
- Zhao, J.; Li, R.; and Jin, Q. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*, 2608–2618.
- Zhao, Z.; Liu, Q.; and Wang, S. 2021. Learning deep global multi-scale and local attention features for facial expression recognition in the Wild. *IEEE Transactions on Image Processing*, 30: 6544–6556.
- Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16816–16825.
- Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2337–2348.
- Zou, S.; Huang, X.; and Shen, X. 2023. Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 5994–6003.