LAMB: Label-Induced Mixed-Level Blending for Multimodal Multi-label Emotion Detection

Shuwei Qian[1,2(✉)], Ming Guo[1,2], Zhicheng Fan[1,2], Mingcai Chen[1,2], and Chongjun Wang[1,2]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
{gm,fanzc,chenmc}@smail.nju.edu.cn, chjwang@nju.edu.cn, qiansw@smail.nju.edu.cn

[2] Department of Computer Science and Technology, Nanjing University, Nanjing, China

Abstract. To better understand complex human emotions, there is growing interest in utilizing heterogeneous sensory data to detect multiple co-occurring emotions. However, existing studies have focused on extracting static information from each modality, while overlooking various interactions within and between modalities. Additionally, the label-to-modality and label-to-label dependencies still lack exploration. In this paper, we propose **LA**bel-induced **M**ixed-level **B**lending (**LAMB**) to address these challenges. Mixed-level blending leverages shallow but manifold self-attention and cross-attention encoders in parallel to model unimodal context dependency and cross-modal interaction simultaneously. This is in contrast to previous works either use one of them or cascade them successively, which ignores the diversity of interaction in multimodal data. LAMB also employs label-induced aggregation to allow different labels to attend to the most relevant blended tokens adaptively using a transformer-based decoder, which facilitates the exploration of label-to-modality dependency. Unlike common low-order strategies in multi-label learning, correlations among multiple labels can be learned by self-attention in label embedding space before being treated as queries. Comprehensive experiments demonstrate the effectiveness of our methods for multimodal multi-label emotion detection.

Keywords: multimodal fusion · multi-label classification · emotion detection

1 Introduction

Detecting emotions plays a vital role in many real-world applications. For example, accurate recognition of emotions is crucial to maintain the high quality of user interaction in some dialogue systems and virtual reality. Moreover, the

H. Gao et al. (Eds.): CollaborateCom 2023, LNICST 562, pp. 20–38, 2024.
https://doi.org/10.1007/978-3-031-54528-3_2

emotional tendencies of the population towards a specific topic on social media can be used to predict large-scale events, such as the general election. With the massive multimodal data accumulated on the Internet, multimodal learning has become a leading approach to this problem. The core challenge of multimodal learning comes from various modality heterogeneity in representation, structure, and semantics.

Since transformer prevails in machine learning, attention mechanisms [23] have become the de-facto paradigm for multimodal representation, alignment, and fusion [1]. Attention mechanisms aggregate values based on the compatibility or similarity between the corresponding keys and the queries. When queries are selected from different modalities, the aggregation patterns of values vary. However, previous works have utilized either self-attention [23] to extract unimodal features or cross-attention [21] to capture cross-modal long-term dependency, ignoring the diverse nature of interaction in multimodal data. In some cases, the context of language modality dominates the labels of the utterance. Taking non-language modality as queries will introduce noise impeding the learning process in such situations. Conversely, utilizing self-attention independently for each modality overlooks the potential interactions among modalities in other cases. Preserving interactions as manifold as possible contributes to a more flexible and powerful fusion.

In addition, most approaches in multimodal emotion detection focus on single-label classification or regression [28,29]. When it comes to the multi-label setting, two additional challenges arise: label-to-modality dependency and label-to-label dependency. Individual labels are influenced by various interactions differently, whereas single-label methods use the same representation for all label classifications. Some emotions, such as anger, often rely more on the coordination of facial expressions and tone of voice than others, which implies different labels depend on different multimodal interactions. Two-stage methods that cascade a multi-label classifier immediately after the unified representation fail to capture the label-to-modality dependency. Besides, some labels tend to co-occur together frequently. For example, people often experience several negative emotions simultaneously when feeling down. This correlation among labels can be used as a priori to guide feature extraction and fusion. Current low-order strategies for multi-label learning discard label correlations [3,5,37] or explore simple pairwise correlations [7,9,10,18] and cannot capture complex correlations among multiple labels.

To tackle these challenges, we propose a method known as **LA**bel-induced **M**ixed-level **B**lending (**LAMB**) for multimodal multi-label emotion detection. Mixed-level blending leverages shallow but diverse self-attention and cross-attention encoders in parallel to model unimodal context dependencies and cross-modal interaction simultaneously. This parallel architecture differs from approaches that stack layers deeply. In this way is retained the diversity of the interactions in multimodal data. Furthermore, LAMB also employs label-induced aggregation to acquire label-specific representations via a transformer-based decoder, which enables the exploration of label-to-modality dependency. Each label embedding selectively attends to the most relevant blended tokens via

cross-attention, adapting to the discriminative information of different emotions. In contrast to common low-order strategies in multi-label learning, correlations among multiple labels are learned by self-attention in label embedding space before being treated as queries to the cross-attention.

Our main contributions can be summarised as follows:

- We propose mixed-level blending, a shallow but manifold architecture with multiple parallel encoders, to preserve diverse interactions of multimodal data, which is essential to a more flexible and powerful fusion.
- We design label-induced aggregation that learns label-specific representations for multiple labels via a transformer-based decoder with self-attention among labels, to explore label-to-modality and label-to-label dependencies.
- Extensive experiments on aligned and unaligned settings demonstrate the superiority of the proposed LAMB and validate the dependencies among modalities and labels.

2 Related Works

2.1 Multimodal Emotion Detection

Relevant works of multimodal emotion detection mainly focus on three challenges: representation, alignment, and fusion.

Representation aims to turn data from different sources with distinct structures into an informative and learnable format to facilitate the coordination of modalities. FDMER [28] and MFSA [29] learn modality-specific and modality-agnostic representations, improving task predictions from the holistic and disentangled views. Self-MM [31] aids the multimodal task via acquiring independent unimodal representations with a self-supervised learning strategy. The vast majority of current works first adopt heterogeneous models to extract unimodal representations from each modality independently, such as LSTM networks for audio and CNN for vision, and leave multimodal representations to the fusion stage.

Alignment refers to identifying the corresponding relationship between subcomponents from one or more modalities of the same instance, which helps to explore the commonality among modalities for a robust prediction. The challenge arises from asynchronous sampling rates and semantic gaps in modalities. CTC network [11] is a typical explicit alignment method that models a probability distribution over all possible label sequences by RNN without pre-segmented data and maximizes the probabilities of the correct one. In order to reuse pre-trained language models, MAG+ [40] proposes an adapter module consisting of cross-modal attention and dynamic gating to align audio and vision with language. PMR [16] introduces a message hub to exchange information across all modalities and progressively reinforces unaligned features. Apart from utilizing alignment modules, there are some works performing alignment by additional objectives. MICA [14] designs a loss function that maximizes mean discrepancy to align the marginal distribution.

Fusion integrates the complementarity of the multimodal data to get a more comprehensive view of the downstream tasks. Previous works have explored various sophisticated fusion mechanisms to achieve this goal, including tensor-based, graph-based, gating-based [19], attention-based methods, etc. TFN [32] and LMF [15] use the outer product between vectors to fuse different levels of multimodal interactions but suffer from large memory costs and computational load. DFG [33] builds a graph where vertices are modal sets and edges denote their inclusion relationship, and fuse modalities via an output vertex connected to all vertices. MISA [12] performs a multi-head self-attention on a concatenation of all the transformed modality vectors to make each vector aware of other modal representations.

There are some noticeable differences between our work and these available works. LAMB exploits multiple label-specific representations for multi-label classification rather than a fixed representation for all labels. Before fusion, it preserves more multimodal dynamics than static modality-specific representations. Besides, our method does not assume alignment between modalities and can be naturally extended to asynchronous multimodal data thanks to the attention mechanism.

2.2 Multi-label Learning

The key challenge of multi-label Learning lies in how to cope with the exponentially growing output space. Although the output space of multi-label data is intimidatingly enormous, labels are not completely separate from each other. Relevant strategies that take advantage of label correlations could be broadly divided into three categories according to the order of label correlations: first-order strategies, second-order strategies, and high-order strategies.

First-order strategies explore labels one-by-one, without considering the relationships among them [3]. The simplest one is a linear classifier with a sigmoid function. Second-order strategies introduce pairwise relations of multiple labels but fail to manage complex real-world applications due to the limitation of their hypothesis [9]. High-order strategies build more complicated label relations to discover effective exploitation of the label correlations [20].

It has been demonstrated in various studies [8,24,38,41] that modeling label correlations can significantly improve classification performance. One example is the use of MLGCN [4], which constructs a label correlation matrix that facilitates the dissemination of information among the nodes in GCN. Other works [13,35,36] have explored label-specific strategies that are efficient for multi-label learning. For instance, LSAN [27] learns the label-specific representation for each document by grabbing the label-related component and computing semantic relations between document words and labels.

In contrast to the above studies, LAMB focuses on multi-label emotion detection in multimodal application scenarios. In addition to label-to-label coexistence, LAMB empowers labels to guide multimodal fusion. Allowing labels to attend diverse multimodal dynamics adaptively is beneficial to exploring label-to-modality dependency, an aspect that remains underdeveloped in other relevant works.

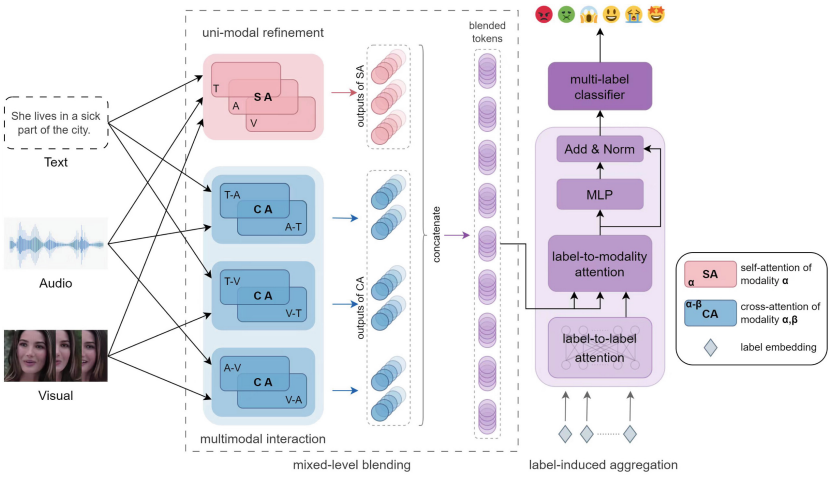

Fig. 1. The overall structure of LAMB. It first encoders different modality tokens by mixed-level blending consisting of unimodal refinement and multimodal interaction modules. All the outputted tokens of both modules are concatenated as blended tokens. Then, these blended tokens and learnable label embeddings are fed into a decoder followed by a linear multi-label classifier.

3 Approach

3.1 Problem Formulation

The goal of multimodal multi-label emotion detection is to predict a set of emotions by utilizing text T, audio A, and visual frames V. We denote a training dataset with N instances as $\mathcal{D} = \{((\mathbf{X}_T, \mathbf{X}_A, \mathbf{X}_V)_i, \mathbf{Y}_i)\}_{i=1}^{N}$. Individual elements in the ith instance $(\mathbf{X}_T, \mathbf{X}_A, \mathbf{X}_V)_i$ are from the original text space $\mathcal{X}^T = \mathbb{R}^{\tau_T \times d_T}$, audio space $\mathcal{X}^A = \mathbb{R}^{\tau_A \times d_A}$, visual space $\mathcal{X}^V = \mathbb{R}^{\tau_V \times d_V}$ respectively, where τ_α and d_α are the maximum sequence length and dimension of modality $\alpha \in \{T, A, V\}$. The label set is $\mathcal{Y} = \{1, 2, \cdots, c\}$, where $c = |\mathcal{Y}|$ is the total number of labels. The task aims to learn a mapping $\mathcal{F} : \mathcal{X}^T \times \mathcal{X}^A \times \mathcal{X}^V \rightarrow 2^{\mathcal{Y}}$ from multiple modalities' joint space to the label space's power set. Since each instance may contain varying numbers of labels, its label set is a subset of $\mathcal{Y}$.

3.2 Overview

As depicted in Fig. 1, LAMB primarily comprises two components: mixed-level blending and label-induced aggregation. In the initial stage, the multimodal data is passed through parallel encoders in mixed-level blending, namely unimodal refinement and multimodal interaction modules, and their outputs are concatenated together as blended tokens. Subsequently, in the label-induced aggregation, trainable label embeddings first attend to each other and then query the

blended tokens that encapsulate diverse multimodal dynamics to obtain label-specific representations. Ultimately, these representations are fed into a linear classifier which outputs a probability distribution over all the labels.

3.3 Mixed-Level Blending

Temporal Convolutions. To bridge the modality gap, we project multimodal data in the original modality space to a unified space $\mathbb{R}^d$, where d is the common dimension for all modalities. Considering the following dot-product attention mechanism utilizes a point-wise similarity to aggregate vectors, it does not give enough attention to the local neighborhood information. LAMB exploits 1D convolutions to alleviate this problem. The following is a description of the process:

$$\hat{\mathbf{X}}_\alpha = \text{Conv1D}(\mathbf{X}_\alpha, \text{kernel}_\alpha), \alpha \in \{T, A, V\} \tag{1}$$

Here, kernel_α is the kernel size of modality α. The attention mechanism which is applied to tokens embedded with local information becomes segment-aware.

Unimodal Refinement Module. Each modality encapsulates distinct and discriminative information, establishing the fundamental basis for prediction. To enhance the predictive capability, it is imperative to fully extract valuable unimodal features from them. Certain pieces of information are contingent upon long-term contextual dependencies, and refining contextual information within each modality facilitates a holistic comprehension of the task at hand. For instance, consider the sentences *'She lives in a sick part of the city. It is full of restaurants.'* If the model fails to grasp the context, it may misinterpret the emotional connotation of 'sick' in the sentence as negative. Conversely, a context-sensitive model would discern that this expression reflects a positive or impressive sentiment in American parlance. To capture the contextual interdependencies within each modality, we utilize self-attention encoders to implement unimodal refinement.

Given queries $\mathbf{Q} \in \mathbb{R}^{n_q \times d_k}$, keys $\mathbf{K} \in \mathbb{R}^{n_k \times d_k}$ and values $\mathbf{V} \in \mathbb{R}^{n_k \times d_v}$, multi-head attention with h heads is employed in accordance with the definition outlined in Transformer [23]:

$$\begin{aligned} \text{head}_i &= \text{softmax}(\frac{\mathbf{Q}\mathbf{W}_i^Q(\mathbf{K}\mathbf{W}_i^K)^T}{\sqrt{\bar{d}_k}})\mathbf{V}\mathbf{W}_i^V \\ \text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{Concat}(\text{head}_1, \cdots, \text{head}_h)\mathbf{W}^O \end{aligned} \tag{2}$$

where $\mathbf{W}_i^Q \in \mathbb{R}^{d_k \times \bar{d}_k}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_k \times \bar{d}_k}$, $\mathbf{W}_i^V \in \mathbb{R}^{d_v \times \bar{d}_v}$, $\mathbf{W}^O \in \mathbb{R}^{h\bar{d}_v \times d_v}$ are the parameter matrices of linear projections. Furthermore, a unimodal refinement module consisting of multi-head attention (MHA), layer normalization (LN), and multilayer perceptron (MLP) of modality α is defined as follows:

$$\begin{aligned} \mathbf{U}_\alpha^0 &= \hat{\mathbf{X}}_\alpha \\ \hat{\mathbf{U}}_\alpha^l &= \text{LN}(\text{MHA}(\mathbf{U}_\alpha^l, \mathbf{U}_\alpha^l, \mathbf{U}_\alpha^l) + \mathbf{U}_\alpha^l) \\ \mathbf{U}_\alpha^{l+1} &= \text{LN}(\text{MLP}(\hat{\mathbf{U}}_\alpha^l) + \hat{\mathbf{U}}_\alpha^l) \end{aligned} \tag{3}$$

where l represents the number of the layer. The simplified notation $\mathbf{U}_\alpha$ refers to the outcomes of the last layer.

Multimodal Interaction Module. Unlike intra-modality context refinement, inter-modality interaction forms new information not involved in a single modality, thus improving predictive performance. For example, *'The movie is very nice'* may convey a negative emotional polarity with an ironic tone. Through the utilization of cross-modal transformers, we model these inter-modality interactions. The cross-modal attention mechanism takes one modality as queries and another modality as key-value pairs. Since queries and key-value pairs originate from distinct modalities, the multimodal interaction module could explore cross-modal long-term dependency without relying on intra-modal context. Concretely, the outputs possess the same length as queries and share identical dimensions with key-value pairs. In other words, the key modality is weighted and averaged based on the query modality rather than itself. Analogous to the unimodal module, a multimodal interaction module that characterizes modality α attending to modality $\beta(\beta \neq \alpha)$ computes the following equations for $l = 0, 1, \cdots$ layers:

$$\begin{aligned}
\mathbf{Z}^0_{\alpha\to\beta} &= \hat{\mathbf{X}}_\alpha, \\
\hat{\mathbf{Z}}^l_{\alpha\to\beta} &= \mathrm{LN}(\mathrm{MHA}(\mathbf{Z}^l_{\alpha\to\beta}, \hat{\mathbf{X}}_\beta, \hat{\mathbf{X}}_\beta) + \mathbf{Z}^l_{\alpha\to\beta}) \\
\mathbf{Z}^{l+1}_{\alpha\to\beta} &= \mathrm{LN}(\mathrm{MLP}(\hat{\mathbf{Z}}^l_{\alpha\to\beta}) + \hat{\mathbf{Z}}^l_{\alpha\to\beta})
\end{aligned} \tag{4}$$

We use $\mathbf{Z}_{\alpha\to\beta}$ to denote the final outputs of the multimodal interaction module.

To preserve the diversity of multimodal dynamics, LAMB concatenates all the tokens outputted from both unimodal refinement and multimodal interaction modules into a new tensor $\ddot{\mathbf{X}}$, namely

$$\begin{aligned}
\ddot{\mathbf{X}} = \mathrm{Concat}(&\mathbf{U}_T, \mathbf{U}_A, \mathbf{U}_V, \mathbf{Z}_{T\to A}, \mathbf{Z}_{T\to V}, \\
&\mathbf{Z}_{A\to T}, \mathbf{Z}_{A\to V}, \mathbf{Z}_{V\to T}, \mathbf{Z}_{V\to A})
\end{aligned} \tag{5}$$

The pertinent discriminative information of all the blended tokens is left for label-induced aggregation to extract in a label-specific manner.

3.4 Label-Induced Aggregation

First-order strategies adopt the label-by-label style, thus neglecting correlations of the other labels. Second-order strategies employ label coexistence, but they are limited to pairwise relations. In contrast, label-induced aggregation explores high-order label-to-label correlations via attention mechanism. Label latent representations enrich the label semantics by allowing each label to attend to all labels, which helps to explore more complicated relations among labels. The initial discrete labels $[1, 2, \cdots, c]$ are first mapped to one-hot vectors $[\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_c]$ where i denotes the location of element *1* in $\mathbf{e}_i$. By looking up a learnable table $\mathbf{W}^E$ that stores embeddings, these one-hot vectors transform into label embeddings $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_c]\mathbf{W}^E$. Then, the embeddings are passed through the

label-to-label self-attention architecture which provides information about all labels to each particular label. After that, label embeddings are treated as queries and fed into the label-to-modality module along with blended tokens as key-value pairs. It is worth mentioning that multimodal tokens $\ddot{\mathbf{X}}$ here are outputs of mixed-level blending. These tokens contain serviceable specific unimodal extraction and complementary multimodal interaction. Unlike relevant works focusing on the CLS token or the last token [21], the label-induced aggregation module retains all tokens. Without token-grained screening, latent representations can be learned and extracted adaptively, thus facilitating the following multiple classifications. The label-to-modality multi-head attention receives label queries alongside modality key-value pairs. Under the guidance of labels, discriminative information is selected to capture potential correlations from modality space to label space. The label-induced aggregation implemented by a transformer-based decoder can be formulated as:

$$\begin{aligned}
\mathbf{D}^0 &= \mathbf{E} \\
\tilde{\mathbf{D}}^l_{L\to L} &= \mathrm{MHA}(\mathbf{D}^l, \mathbf{D}^l, \mathbf{D}^l) \\
\hat{\mathbf{D}}^l_{L\to M} &= \mathrm{MHA}(\tilde{\mathbf{D}}^l_{L\to L}, \ddot{\mathbf{X}}, \ddot{\mathbf{X}}) \\
\mathbf{D}^{l+1} &= \mathrm{LN}(\mathrm{MLP}(\hat{\mathbf{D}}^l_{L\to M}) + \hat{\mathbf{D}}^l_{L\to M})
\end{aligned} \tag{6}$$

where $\mathbf{D}^0$ refers to the initial state of the first layer in the label-induced aggregation decoder. Similarly, $\mathbf{D}^l$ and $\mathbf{D}^{l+1}$ denote the outputs of the lth and $(l+1)$th layers in the decoder. $\tilde{\mathbf{D}}^l_{L\to L}$ and $\hat{\mathbf{D}}^l_{L\to M}$ are two intermediate states. The self-attention in computing $\tilde{\mathbf{D}}^l_{L\to L}$ and the cross-attention in computing $\hat{\mathbf{D}}^l_{L\to M}$ explore the label-to-label and label-to-modality dependencies respectively.

3.5 Multi-label Classifier

After obtaining the label-specific representations $\mathbf{D}_{L\to M} \in \mathbb{R}^{N\times c\times d}$, LAMB feeds them into a linear classifier with weight matrix $\mathbf{W}^f \in \mathbb{R}^{d\times 1}$, bias $b \in \mathbb{R}$, and sigmoid function to infer label sets. The predicted probabilities of each label are written as:

$$\hat{\mathbf{Y}} = f(\mathbf{D}_{L\to M}) = \mathrm{sigmoid}(\mathbf{D}_{L\to M}\mathbf{W}^f + b^f) \tag{7}$$

Finally, we use binary Cross-entropy loss which is a classical optimization objective for multi-label learning to guide our model globally for classification. Formally, it can be calculated as:

$$\mathrm{BCELoss}(\mathbf{Y}, \hat{\mathbf{Y}}) = -\frac{1}{N}\sum_{i=1}^{N}\Big[\mathbf{Y}_i \log(\hat{\mathbf{Y}}_i) + (1-\mathbf{Y}_i)\log(1-\hat{\mathbf{Y}}_i)\Big] \tag{8}$$

4 Experiments

4.1 Experimental Settings

Dataset. The CMU-MOSEI [33] dataset is a widely used benchmark for multimodal sentiment analysis and emotion recognition. It comprises 3,229 full-length videos of 1,000 speakers, divided into 22,856 video segments at the utterance level. Each video segment consists of three modalities: text, audio, and visual, and covers six emotion categories: anger, disgust, fear, happiness, sadness, and surprise. To extract features from the visual modality, FACET [2] employs a facial expression recognition system that generates 35-dimensional features from video frames. For the audio modality, COVAREP [6] extracts 74-dimensional features from acoustic signals. Lastly, GloVe [17] provides 300-dimensional features from the video transcripts for the text modality.

Evaluation Criteria. Since multimodal multi-label emotion detection is a classification task, we report the performance of all approaches with four classical metrics and a typical multi-label loss function, i.e., accuracy (Acc), precision (P), recall (R), micro-F1 (F1), and Hamming loss (HL). Superior performance is indicated by higher accuracy, precision, recall, and micro-F1 scores, while lower values of Hamming loss are desirable outcomes.

Implement Details. We implement the LAMB model using Pytorch and conduct all evaluations on an NVIDIA RTX 3090 GPU. The model is trained by Adam optimizer with default parameters. In unimodal refinement modules, multimodal interaction modules, and the label-induced aggregation decoder, the number of layers is set to 1, as LAMB prioritizes diversity over depth. The attention heads in all attention layers are fixed to 5. The dimension of the unified modality space, as well as the hidden sizes of the transformer encoders and the decoder, have been configured to 30. Notably, the hidden size of the intermediate layer within the transformer block's multilayer perceptron is set to 120 for the encoders and 256 for the decoder.

Additionally, we employ label smoothing to prevent overfitting and clip the gradients to the maximum norm of 0.8. During training, we train each model for a fixed number of epochs 50 with an early-stopping strategy and schedule the learning rate from 1e-3 by its performance on the validation set. After the training process, we select the model with the highest accuracy on the validation set as our final model and evaluate its performance on the test set.

4.2 Baselines

Based on the degree of modality contribution and label correlation participation, the baselines in our study can be categorized into four groups. Firstly, classical methods transform multi-label learning problems into other sophisticated learning scenarios. Regardless of modality heterogeneity, the multimodal inputs are concatenated as new inputs. **BR** [3] decomposes the multi-label task into

independent binary classification problems, ignoring the correlations between labels. **LP** [22] converts the original multi-label set into small random subsets and tackles the modified single-label classification problem. **CC** [20] transforms the multi-label learning problem into a chain of binary classification problems, considering high-order label correlations. Secondly, text-based or image-based algorithms solely utilize information from the text or image modality. **SGM** [30] treats multi-label classification as a sequence generation problem to exploit label relations for text data. **LSAN** [27] constructs label-specific document representation by simultaneously using document content and label text. **ML-GCN** [4] proposes a graph convolutional network-based model to capture label dependencies for multi-label image recognition. Thirdly, this group of baselines mainly focuses on solving multimodal issues but lacks the utilization of label correlation. **DFG** [33] analyzes the mechanism of modality interaction in sentiment analysis and emotion recognition by taking advantage of the interpretable dynamic fusion graph algorithm. **RAVEN** [25] investigates the fine-grained structure of nonverbal subword sequences and constructs multimodal-shifted word representations to dynamically capture changes in non-linguistic context. **MulT** [21] introduces a cross-modal attention mechanism to provide a latent cross-modal adaptation for multimodal fusion, while capturing long-range contingencies. **SIMM** [26] develops shared subspace and extracts view-specific information to strengthen communication between views while preserving individual-specific

Table 1. Performance comparison on aligned and unaligned settings. The best results are in bold. LAMB outperforms other state-of-the-art approaches, indicating the effectiveness of our method.

Approaches	Aligned				Unaligned			
	Acc ↑	P ↑	R ↑	F1 ↑	Acc ↑	P ↑	R ↑	F1 ↑
BR [3]	0.222	0.309	0.515	0.386	0.233	0.321	0.545	0.404
LP [22]	0.159	0.231	0.377	0.286	0.185	0.252	0.427	0.317
CC [20]	0.225	0.306	0.523	0.386	0.235	0.320	0.550	0.404
SGM [30]	0.455	0.595	0.467	0.523	0.449	0.584	0.476	0.524
LSAN [27]	0.393	0.550	0.459	0.501	0.403	0.582	0.460	0.514
ML-GCN [4]	0.411	0.546	0.476	0.509	0.437	0.573	0.482	0.524
DFG [33]	0.396	0.595	0.457	0.517	0.386	0.534	0.456	0.494
RAVEN [25]	0.416	0.588	0.461	0.517	0.403	0.633	0.429	0.511
MulT [21]	0.445	0.619	0.465	0.531	0.423	0.636	0.445	0.523
SIMM [26]	0.432	0.561	0.495	0.525	0.418	0.482	0.486	0.484
MISA [12]	0.430	0.453	**0.582**	0.509	0.398	0.371	**0.571**	0.450
HHMPN [34]	0.459	0.602	0.496	0.556	0.434	0.591	0.476	0.528
TAILOR [39]	0.488	0.641	0.512	0.569	0.460	0.639	0.452	0.529
LAMB (Ours)	**0.490**	**0.643**	0.517	**0.573**	**0.463**	**0.656**	0.454	**0.536**

characteristics. **MISA** [12] obtains modality-invariant and modality-specific features that are fused to estimate affective states by considering the holistic view of the multimodal data. Fourthly, multimodal multi-label approaches validate the significance of exploring modalities and labels comprehensively. **HHMPN** [34] effectively handles both complete and partial time series data. The feature-to-label, label-to-label and modality-to-label dependencies are modeled simultaneously by means of the graph message passing. **TAILOR** [39] extracts private and common representations adversarially and leverages label semantics to construct label-specific representations for multimodal fusion at different granularity.

5 Results and Analysis

5.1 Comparison Experiment

Table 1 shows the performance of representative approaches to multimodal multi-label emotion detection on CMU-MOSEI in both aligned and unaligned settings. Due to the heterogeneity of modalities, different modalities are obtained at different sampling rates. As a result, the three modalities possess varying sequence lengths for the same utterance in the unaligned setting. In the aligned setting, preprocessed video and audio data are aligned with the words in the text, ensuring that all modalities have the same sequence length. For methods incapable of directly handling non-aligned data, we report the results of their modified version that contain an additional CTC module [11] and corresponding training loss in the unaligned setting. Based on the comparison results, there are some observations:

1. CC achieves the best performance among the three classical multi-label methods, validating the utilization of label correlation is beneficial for multi-label learning.
2. The classical multi-label approaches BR, LP, and CC show much worse performance than the text-based or image-based algorithms SGM, LSAN, and ML-GCN, which indicates that methods employing fully exploited unimodal features are comparable with classical multi-label approaches.
3. Most of the multimodal baselines surpass text-based or image-based. Specifically, MulT outperforms SGM except for accuracy, ML-GCN except for recall, and beats LSAN through all metrics in the aligned setting, demonstrating the necessity of exploiting intra-modal and inter-modal dynamics.
4. Multimodal multi-label methods, such as TAILOR, display even better outcomes than the approaches mentioned above, suggesting the effectiveness of simultaneously leveraging modalities and labels.
5. In both settings, our proposed LAMB demonstrates superior performance in addressing multi-label emotion detection across three metrics, albeit with suboptimal recall results. This is attributed to our approach effectively exploring label-to-label and label-to-modality dependencies while preserving the diversity of multimodal interactions.

5.2 Ablation Study

With the purpose of figuring out the role of individual components in LAMB, we further conduct ablation experiments on CMU-MOSEI. By selectively removing different parts while keeping the basic structure, the performance variations of specific components reveal their impact. Table 2 presents the evaluation results. We report four typical multi-label evaluation metrics: Accuracy, Hamming loss, Recall, and Micro-F1.

Table 2. Ablation study on aligned CMU-MOSEI dataset. The best results are in bold. "w/o" denotes removing the component.

Approches	Acc ↑	R ↑	F1 ↑	HL ↓
only text	0.456	0.452	0.544	0.260
only audio	0.418	0.435	0.498	0.266
only visual	0.429	0.413	0.502	0.287
only text-audio	0.458	0.466	0.546	0.248
only text-visual	0.460	0.456	0.551	0.257
only audio-text	0.472	0.477	0.553	0.241
only audio-visual	0.445	0.452	0.518	0.256
only visual-text	0.481	0.489	0.556	0.233
only visual-audio	0.436	0.437	0.507	0.266
only unimodal	0.482	0.512	0.566	0.220
only multimodal	0.480	0.496	0.563	0.229
w/o decoder	0.474	0.479	0.565	0.241
w/o label correlations	0.478	0.512	0.552	0.220
w/o label embeddings	0.479	0.491	0.561	0.232
LAMB (Ours)	**0.490**	**0.517**	**0.573**	**0.217**

Role of Individual Modalities. Avoiding the mutual influence among modalities, three variants using only a single modality as input are scrutinized. *'only text'*, *'only audio'*, and *'only visual'* in Table 2 denote experiments to explore the role of individual modalities with models that only use a single unimodal refinement module and do not use multimodal interaction modules. For instance, *'only text'* represents the model only using a text unimodal refinement module without any multimodal interaction modules followed by a label-induced aggregation directly. As is exhibited, the result obtained by using only one modality input is the worst among all the results, which fully reflects the importance of utilizing information from multiple modalities. Multimodal complementarity has a powerful impetus on emotion recognition tasks.

Role of Paired Modality Interaction. We perform experiments to explore the role of paired modality interaction with models that only use one multimodal interaction module and do not use unimodal refinement modules. As two arbitrary modalities serve as either queries or key-value pairs in the cross-attention mechanism, this results in six different variants. For example, *'only text-audio'* denotes the model that takes text as queries and audio as key-value pairs. As illustrated in Table 2, our LAMB surpasses all the variants. We can conclude that preserving the diversity of different multimodal interactions is crucial to a more flexible and powerful fusion.

Role of Mixed-Level Blending. To explore the role of unimodal refinement and multimodal interaction in mixed-level blending, we conduct experiments denoted by *'only unimodal'* and *'only multimodal'*. *'only unimodal'* represents the variant where three unimodal refinement modules of text, audio, and visual modalities are followed by a label-induced aggregation without any multimodal interaction modules, while *'only multimodal'* represents the model with all six multimodal interaction modules followed by label-induced aggregation directly. The poor performance of these two variants implies the effectiveness of exploiting the intra-modality context in a single modality and exploring long-term inter-modality dependency.

Role of Label-Induced Aggregation. Furthermore, label induction is under evaluation. The first variant, as is demonstrated in the *'w/o decoder'* of Table 2, corresponds to the whole decoder being removed. Instead, the first element of the tokens from self-attention and cross-modal attention are concatenated directly. This variant is the worst result among the three in this label impact detection section, signifying the indispensability of the whole label-induced aggregation. In other words, only taking advantage of modality information is inadequate. In order to probe the necessity of label correlations, we replace the regular attention masks with an identity matrix in the label-to-label self-attention so that each label can only attend to itself without any correlations. *'w/o label correlations'* displays the performance of this variant which treats each label independently. By comparison, we can conclude that the attention mechanism indeed makes contributions to label-to-label correlations. For the sake of proving the mutual effect between labels and modalities, we use identical rather than label-specific embeddings to fuse modalities for all the labels as the third variant. As is shown in *'w/o label embeddings'*, the defective consequence of eliminating label-specific guidance shows the importance of label-to-modality dependency.

As is demonstrated in Table 2, all measurements decline consistently, regardless of the component being removed. It reveals that each part plays an irreplaceable role in achieving the remarkable performance of LAMB. The complete version encompassing all elements is better than other variants, underscoring the mutual enhancement among all components.

5.3 Visualization

Analysis on Label-to-modality Dependency. To further investigate the label-to-modality dependency, we select the tokens that receive the highest five attention values across all the heads for each emotion. These tokens contain the most discriminative information for the corresponding labels. Then, we count the number of these tokens each encoder generates and plot the distribution in the histogram. The higher the number is, the more important this kind of multimodal dynamics is for the specific emotion. In Fig. 2, 't', 'a', and 'v' denote unimodal refinement modules from text, audio, and visual respectively. Similarly, 'ta' refers to the multimodal interaction module with text as queries, audio as key-value pairs, and so on. The highest bars of different labels have different colors, demonstrating that emotions tend to depend on different multimodal

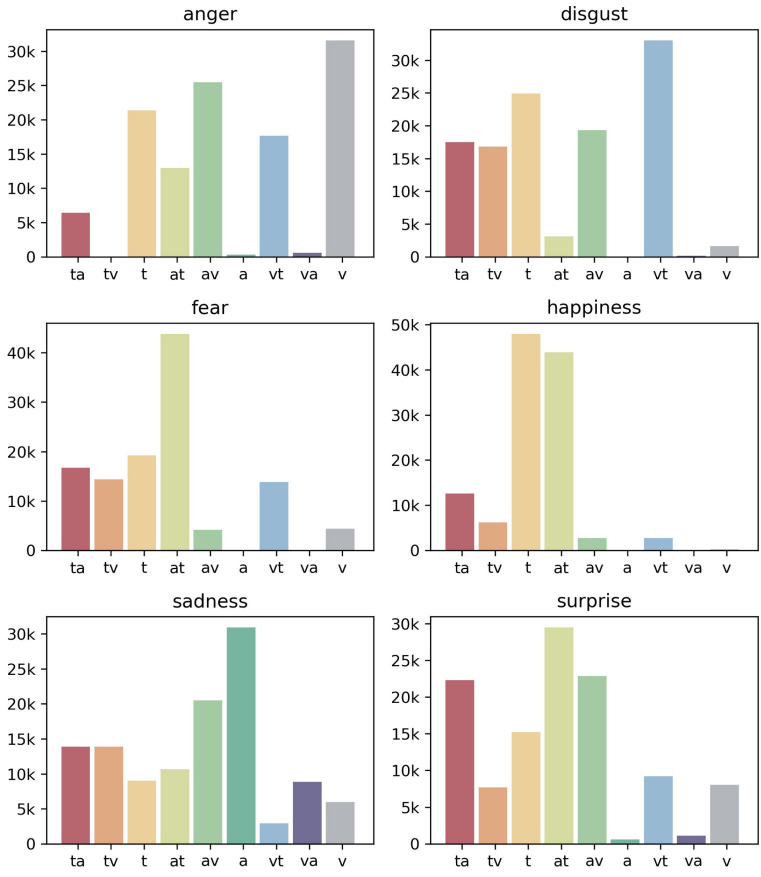

Fig. 2. Histogram of the encoders from which the top-5 attended tokens are over six emotions. The highest bar differs in each chart, which indicates labels have distinct associations with different multimodal dynamics and LAMB adaptively learns it.

dynamics. For example, facial expression is vital to recognize anger, while audio-text interactions are more crucial for fear. Thanks to label-induced aggregation, LAMB explores the label-to-modality dependency adaptively.

Analysis on Label-to-Label Dependency. In order to delve into the label-to-label dependency, we extract all the attention matrices in the label self-attention of the decoder and visualize them in heatmaps. As shown in Fig. 3, (a) - (e) are the dependence among different labels learned by five attention heads, and (f) is the outcome of using the identity matrix. The label-to-label self-attention models complicated label correlations compared to utilizing the identity matrix. Furthermore, different attention heads learn diverse label co-existence to enrich the semantic information from various perspectives. For instance, anger has a

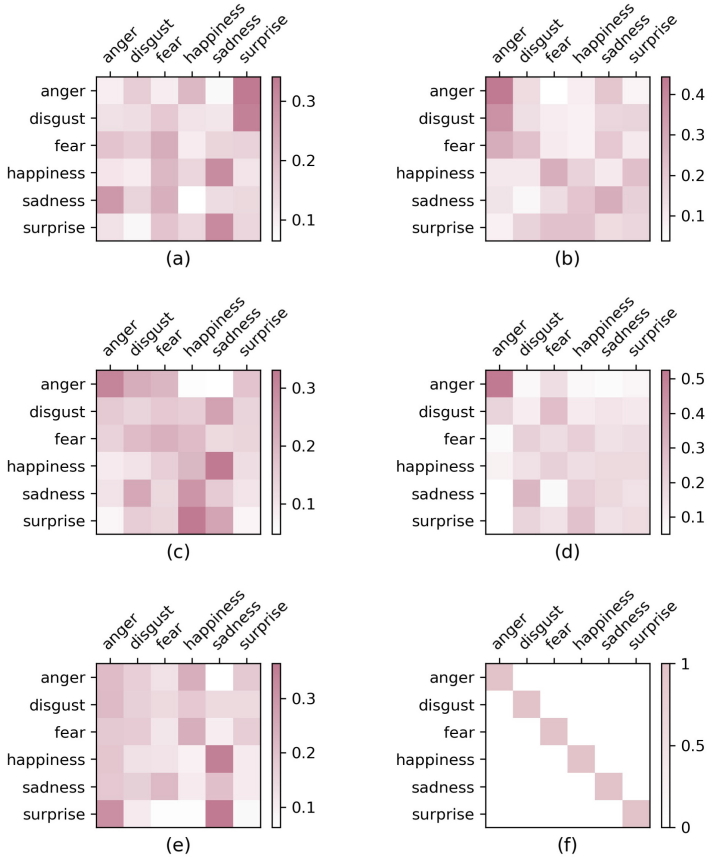

Fig. 3. Visualization of label attention matrices. The difference between (a) - (e) and the identity matrix (f) demonstrates LAMB can capture complex label correlations.

higher correlation with surprise in the first head (a), while it pays more attention to itself in other heads.

6 Conclusion

In this paper, we propose LAMB, a novel approach for multimodal multi-label emotion detection that addresses the inefficiency of static modality information extraction, as well as the limited exploration of label-to-modality and label-to-label dependencies. LAMB mainly consists of mixed-level blending and label-induced aggregation. Mixed-level blending involves multiple parallel encoders to preserve modality interaction diversity for better fusion. Label-induced aggregation obtains label-specific representation by employing learnable label embeddings to query blended tokens. It allows the model to capture label-to-modality and label-to-label dependencies, which are essential for accurate emotion detection.

Experimental evaluations and analysis of aligned and unaligned data certify the effectiveness of our proposed LAMB. Individual component of LAMB plays an indispensable role and coordinates with each other. Label correlations are explored by self-attention in label embedding space. Under the induction of labels, tokens from different encoders provide unequal discriminative information for different emotions, validating the dependencies between modalities and labels.

Emotion detection models often lack interpretability, making it difficult to understand the reasoning behind their predictions. Future work could delve into techniques for interpreting and explaining the decisions made by the LAMB framework. This could enable users to gain insights into the model's decision-making process and enhance trust and transparency in its applications.

Acknowledgments. This paper is supported by the National Natural Science Foundation of China (Grant No. 62192783, 62376117), the Collaborative Innovation Center of Novel Software Technology and Industrialization at Nanjing University.

References

1. Baltrusaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 423–443 (2019)
2. Baltrusaitis, T., Robinson, P., Morency, L.P.: OpenFace: an open source facial behavior analysis toolkit. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 1–10 (2016)
3. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern Recogn. **37**, 1757–1771 (2004)
4. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5177–5186 (2019)
5. Clare, A., King, R.D.: Knowledge discovery in multi-label phenotype data. In: Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery, pp. 42–53 (2001)

6. Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: COVAREP - A collaborative voice analysis repository for speech technologies. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 960–964 (2014)
7. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: Proceedings of the Conference on Neural Information Processing Systems, pp. 681–687 (2001)
8. Feng, L., An, B., He, S.: Collaboration based multi-label learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3550–3557 (2019)
9. Fürnkranz, J., Hüllermeier, E., Mencía, E.L., Brinker, K.: Multilabel classification via calibrated label ranking. Mach. Learn. **73**, 133–153 (2008)
10. Ghamrawi, N., McCallum, A.: Collective multi-label classification. In: Proceedings of the ACM International Conference on Information and Knowledge Management, pp. 195–200 (2005)
11. Graves, A., Fernández, S., Gomez, F.J., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the International Conference on Machine Learning, pp. 369–376 (2006)
12. Hazarika, D., Zimmermann, R., Poria, S.: MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In: Proceedings of the ACM International Conference on Multimedia, pp. 1122–1131 (2020)
13. Huang, J., Li, G., Huang, Q., Wu, X.: Learning label-specific features and class-dependent labels for multi-label classification. IEEE Trans. Knowl. Data Eng. **28**, 3309–3323 (2016)
14. Liang, T., Lin, G., Feng, L., Zhang, Y., Lv, F.: Attention is not Enough: mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8128–8136 (2021)
15. Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Zadeh, A., Morency, L.P.: Efficient low-rank multimodal fusion with modality-specific factors. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 2247–2256 (2018)
16. Lv, F., Chen, X., Huang, Y., Duan, L., Lin, G.: Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2554–2562 (2021)
17. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543 (2014)
18. Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Mei, T., Zhang, H.J.: Correlative multi-label video annotation. In: Proceedings of the ACM International Conference on Multimedia, pp. 17–26 (2007)
19. Rahman, W., et al.: Integrating multimodal information in large pretrained transformers. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 2359–2369 (2020)
20. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Mach. Learn. **85**, 333–359 (2011)
21. Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 6558–6569 (2019)

22. Tsoumakas, G., Katakis, I.: Multi-label classification: an overview. Int. J. Data Warehouse. Min. **3**, 1–13 (2007)
23. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
24. Wang, H., et al.: Collaboration based multi-label propagation for fraud detection. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 2477–2483 (2020)
25. Wang, Y., Shen, Y., Liu, Z., Liang, P.P., Zadeh, A., Morency, L.P.: Words Can Shift: dynamically adjusting word representations using nonverbal behaviors. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7216–7223 (2019)
26. Wu, X., et al.: Multi-View Multi-label learning with view-specific information extraction. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 3884–3890 (2019)
27. Xiao, L., Huang, X., Chen, B., Jing, L.: Label-specific document representation for multi-label text classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 466–475 (2019)
28. Yang, D., Huang, S., Kuang, H., Du, Y., Zhang, L.: Disentangled representation learning for multimodal emotion recognition. In: Proceedings of the ACM International Conference on Multimedia, pp. 1642–1651 (2022)
29. Yang, D., Kuang, H., Huang, S., Zhang, L.: Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences. In: Proceedings of the ACM International Conference on Multimedia, pp. 1708–1717 (2022)
30. Yang, P., Sun, X., Li, W., Ma, S., Wu, W., Wang, H.: SGM: sequence generation model for multi-label classification. In: Proceedings of the International Conference on Computational Linguistics, pp. 3915–3926 (2018)
31. Yu, W., Xu, H., Yuan, Z., Wu, J.: learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 10790–10797 (2021)
32. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1103–1114 (2017)
33. Zadeh, A., Liang, P.P., Poria, S., Cambria, E., Morency, L.P.: Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 2236–2246 (2018)
34. Zhang, D., et al.: Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 14338–14346 (2021)
35. Zhang, M.L., Fang, J.P., Wang, Y.B.: BiLabel-specific features for multi-label classification. ACM Trans. Knowl. Discov. Data **16**, 1–23 (2022)
36. Zhang, M.L., Wu, L.: Lift: multi-label learning with label-specific features. IEEE Trans. Knowl. Data Eng. **37**, 107–120 (2015)
37. Zhang, M.L., Zhou, Z.H.: ML-KNN: A lazy learning approach to multi-label learning. Pattern Recogn. **40**, 2038–2048 (2007)
38. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE Trans. Knowl. Data Eng. **26**, 1819–1837 (2014)
39. Zhang, Y., Chen, M., Shen, J., Wang, C.: Tailor versatile multi-modal learning for multi-label emotion recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 9100–9108 (2022)

40. Zhao, X., Chen, Y., Li, W., Gao, L., Tang, B.: MAG+: an extended multimodal adaptation gate for multimodal sentiment analysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4753–4757 (2022)
41. Zhu, Y., Kwok, J.T., Zhou, Z.H.: Multi-label learning with global and local label correlation. IEEE Trans. Knowl. Data Eng. **30**, 1081–1094 (2018)