Efficient Multimodal Transformer With Dual-Level Feature Restoration for Robust Multimodal Sentiment Analysis

Licai Sun, Zheng Lian, Bin Liu, *Member, IEEE*, and Jianhua Tao, *Senior Member, IEEE*

***Abstract*—With the proliferation of user-generated online videos, Multimodal Sentiment Analysis (MSA) has attracted increasing attention recently. Despite significant progress, there are still two major challenges on the way towards robust MSA: 1) inefficiency when modeling cross-modal interactions in unaligned multimodal data; and 2) vulnerability to random modality feature missing which typically occurs in realistic settings. In this paper, we propose a generic and unified framework to address them, named Efficient Multimodal Transformer with Dual-Level Feature Restoration (EMT-DLFR). Concretely, EMT employs utterance-level representations from each modality as the global multimodal context to interact with local unimodal features and mutually promote each other. It not only avoids the quadratic scaling cost of previous local-local cross-modal interaction methods but also leads to better performance. To improve model robustness in the incomplete modality setting, on the one hand, DLFR performs low-level feature reconstruction to implicitly encourage the model to learn semantic information from incomplete data. On the other hand, it innovatively regards complete and incomplete data as two different views of one sample and utilizes siamese representation learning to explicitly attract their high-level representations. Comprehensive experiments on three popular datasets demonstrate that our method achieves superior performance in both complete and incomplete modality settings.**

***Index Terms*—Multimodal sentiment analysis, unaligned and incomplete data, efficient multimodal Transformer, dual-level feature restoration, robustness.**

I. Introduction

MULTIMODAL Sentiment Analysis (MSA), which leverages multimodal signals to achieve an affective understanding of user-generated video, has become an active research area due to its wide applications in marketing management [1], [2], social media analysis [3], [4], and human-computer interaction [5], [6], to name a few. It mainly involves sequential data of three common modalities, i.e., audio (acoustic behaviors), vision (facial expressions), and text (spoken words). These different types of data provide us with abundant information to make a thorough understanding of human sentiment. Nevertheless, it remains challenging to efficiently fuse the heterogeneous sequential features in practical applications. The first issue is that multimodal sequences usually exhibit *unaligned* nature as different modalities typically have variable sampling rates. In addition, they often suffer from random modality feature missing (i.e., *incomplete*) due to many inevitable factors in real-world scenarios. For instance, the speech may be temporarily corrupted by background noise or sensor failure. The speaker's face could occasionally miss because of occlusion and motion. Some spoken words are probably unavailable owing to automatic speech recognition errors.

The earlier studies address the *unaligned* issue by manually performing forced word-level alignment before model training [7], [8], [9], [10], [11]. However, the manual alignment process requires domain expert knowledge and is not always feasible in the real-world deployment of MSA models. Recently, Multimodal Transformer (MulT) [12] has been proposed to directly model cross-modal correlations in unaligned multimodal sequences. It utilizes directional pairwise cross-modal attention to attend to dense (i.e., *local-local*) interactions across distinct time steps between two involved modalities. Although MulT can address the unaligned issue, it is not efficient to conduct multimodal fusion in a pairwise manner. Therefore, Lv et al. [13] propose the Progressive Modality Reinforcement (PMR), which introduces a message hub to communicate with each modality. The message hub can send common messages to each modality and it can also collect information from them. In this way, PMR avoids the inefficient pairwise communication of two modalities in MulT. Unfortunately, both PMR and MulT suffer from quadratic computational complexity over the involved modalities, as they focus on modeling the *local-local* cross-modal dependencies across all modalities (see details in Section III-C).

Moreover, random modality feature missing which often occurs in realistic settings exacerbates the problem of fusing unaligned multimodal sequences. A simple method to tackle the *incomplete* issue is to perform zero, average, or nearest-neighbor imputation during model inference. However, there is a large

Manuscript received 16 August 2022; revised 19 February 2023; accepted 29 April 2023. Date of publication 10 May 2023; date of current version 1 March 2024. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 61831022, 62276259, 62201572, and U21B2010, in part by Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park under Grant Z211100004821013, in part by the Open Research Projects of Zhejiang Lab under Grant 2021KH0AB06, and in part by CCF-Baidu Open Fund under Grant OF2022025. Recommended for acceptance by C. Lee. *(Corresponding Authors: Bin Liu and Jianhua Tao.)*

Licai Sun is with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with National Laboratory of attern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: sunlicai2019@ia.ac.cn).

Zheng Lian and Bin Liu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: lianzheng2016@ia.ac.cn; liubin@nlpr.ia.ac.cn).

Jianhua Tao is with the Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: jhtao@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TAFFC.2023.3274829

domain gap between the ground truth and the imputed one. Thus, the model performance typically degrades severely using these naive imputations. Recently, several modality translation-based methods [11], [14], [15] have been proposed to learn robust joint multimodal representations that retain maximal information from all modalities. While appealing, these methods are developed to handle the entire loss of one or more modalities which is less likely to happen in practice. Besides, most of them are only applicable to aligned multimodal inputs. More recently, Yuan et al. [16] introduce the Transformer-based Feature Reconstruction Network (TFR-Net) to cope with random modality feature missing. It expects the model to implicitly learn semantic features from incomplete multimodal sequences by reconstructing the missing parts. Though achieving promising results, TFR-Net has the risk of learning trivial solutions as the model may find a shortcut instead of inferring semantics to solve the reconstruction task [17]. Besides, it builds on MulT and thus also has a quadratic scaling cost with the involved modalities.

To efficiently fuse unaligned multimodal sequences in both complete and incomplete modality settings, we propose a generic and unified framework in this paper, named Efficient Multimodal Transformer with Dual-Level Feature Restoration (EMT-DLFR). In contrast to MulT and PMR, the Efficient Multimodal Transformer (EMT) explores less dense (i.e., *global-local*) cross-modal interactions, which is mainly motivated by the observation that a large amount of redundancy exists in unaligned multimodal sequences (especially for the audio and video modalities which have high sampling rates). Inspired by the bottleneck mechanism in Transformers [18], [19], [20], EMT utilizes utterance-level representations from each modality as the *global* multimodal context to interact with *local* unimodal features. On the one side, local unimodal features can be efficiently reinforced by the global multimodal context via cross-modal attention. In turn, the global multimodal context can update itself by extracting useful information from local unimodal features through symmetric cross-modal attention. By stacking multiple layers, the global multimodal context and local unimodal features can mutually promote each other and refine themselves progressively. Thanks to the introduction of the global multimodal context, EMT not only practically has linear computational complexity over the involved modalities but also leads to performance gains. Furthermore, we introduce hierarchical parameter sharing for EMT to increase parameter efficiency and ease model training.

To promote model robustness to random modality feature missing, we randomly mask the input feature sequences of complete data to mimic real-world scenarios and utilize the Dual-Level Feature Restoration (DLFR) built upon EMT to achieve robust representation learning from incomplete multimodal data. On the one hand, DLFR follows TFR-Net and tries to reconstruct the missing parts by exploiting available intra- and inter-modal clues using multiple stacked layers in EMT, i.e, the Low-Level Feature Restoration (LLFR), or more specifically, *low-level* feature reconstruction. In this way, LLFR *implicitly* encourages the model to learn semantic information from incomplete multimodal sequences. On the other hand, inspired by recent advances in self-supervised representation learning [21], [22], [23], we originally regard incomplete and complete sequences as two different views of one sample, and utilize siamese representation learning to *explicitly* attract high-level representations of incomplete and complete views in the latent space, i.e., High-Level Feature Restoration (HLFR), or more specifically, *high-level* feature attraction. Compare with LLFR, HLFR is more direct and thus more effective. Nevertheless, they are complementary to each other. Therefore, the combination of LLFR and HLFR can be a unified framework for robust MSA.

To verify the effectiveness of the proposed method, we conduct comprehensive experiments on three widely used MSA benchmark datasets, including CMU-MOSI [24], CMU-MOSEI [25], and CH-SIMS [26]. The results show that our proposed method outperforms previous state-of-the-art methods in both incomplete and complete modality settings. To summarize, the main contributions of this paper are as follows:

- We propose EMT, an Efficient Multimodal Transformer to achieve effective and efficient fusion of unaligned multimodal data. It not only avoids the quadratic scaling cost of previous dense cross-modal interaction methods but also achieves better performance than them.
- We propose DLFR, a Dual-Level Feature Restoration method to improve model robustness to random modality feature missing which typically occurs in real-world scenarios. It combines both implicit low-level feature reconstruction and explicit high-level feature attraction to realize robust representation learning from incomplete multimodal data.
- Extensive experiments on three popular MSA benchmark datasets demonstrate that EMT and EMT-DLFR achieve state-of-the-art performance in the complete and incomplete modality settings, respectively. [1]

II. Related Work

A. MSA in the Complete Modality Setting

Most works in MSA assume a complete modality setting and they are centered around developing various methods to fuse heterogeneous features from different modalities. Generally, they conduct multimodal fusion at two different levels: 1) utterance level, and 2) element level. The utterance-level fusion methods mainly include simple concatenation [27], [28], [29], [30], attention [7], [31], [32], and tensor-based fusion [33], [34], [35], [36]. Although these methods obtain promising results, they ignore the fine-grained cross-modal interactions. Therefore, lots of methods have been proposed to perform element-level fusion on manually aligned multimodal sequences, including recurrent methods [8], [9], [37], attention-based methods [10], [38], and multimodal-aware word embeddings [39], [40]. Unfortunately, the manual alignment process requires domain expert knowledge and is not always feasible in real-world applications. Motivated by the great success of Transformer [41] in various fields of deep learning, Transformer-based methods [12], [13], [16], [42], [43], [44], [45], [46] have attracted increasing attention

[1] The code will be available at https://github.com/sunlicai/EMT-DLFR.

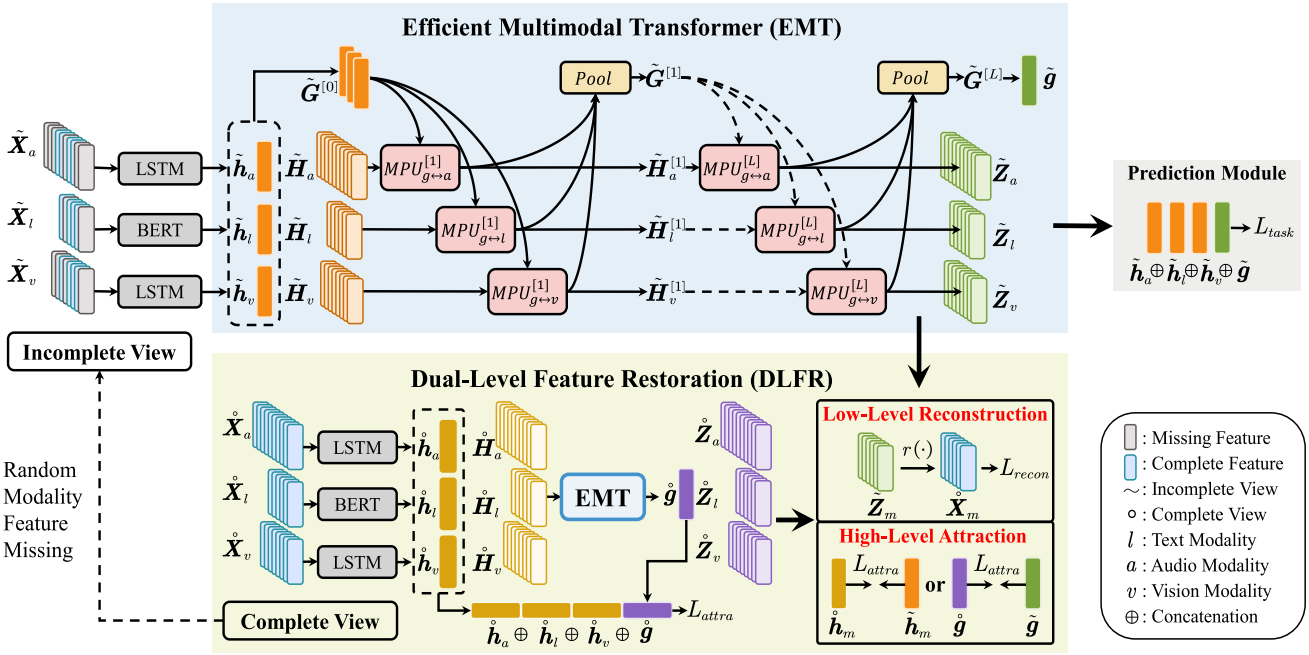


Fig. 1. The overall architecture of EMT-DLFR, which mainly consists of the Efficient Multimodal Transformer (EMT) and Dual-Level Feature Restoration (DLFR) for efficient and robust MSA. To achieve efficient multimodal fusion, EMT utilizes utterance-level representations from each modality as the global multimodal context G to interact with local unimodal features H_m ($m \in \{l, a, v\}$) through the Mutual Promotion Unit (MPU). Based on EMT, DLFR employs both explicit high-level feature attraction and implicit low-level feature reconstruction to guide robust representation learning from incomplete multimodal sequences. Best viewed in color.

in recent years as they can directly perform multimodal fusion on unaligned multimodal sequences. For instance, MulT [12] utilizes directional pairwise cross-modal attention to attend to dense interactions between different modalities. PMR [13] further introduces a message hub to explore multi-way interactions in a single cross-modal attention module. TFR-Net [16] adds an extra intra-modal Transformer in parallel with the cross-modal Transformer in MulT to achieve simultaneous modeling of intra-modal and cross-modal interactions. Although achieving encouraging results, they all have quadratic computational complexity over the involved modalities. In contrast, due to the introduction of the global multimodal context, our proposed method enjoys the linear scaling cost in practice and even has better performance.

B. MSA in the Incomplete Modality Setting

Compared with numerous methods in the above setting, there are only a few studies focusing on improving model robustness in the incomplete setting, despite its critical role in enabling reliable deployment in the wild. Parthasarathy et al. [47] propose a strategy to cope with incomplete sequences by randomly ablating input features during training. Similarly, Hazarika et al. [48] utilize modality-perturbation (i.e., removing or masking modalities) training to reduce the sensitivity of the model to the missing language modality. Liang et al. [49] present a tensor rank regularization method to learn multimodal representations from aligned imperfect time series data. Pham et al. [11] utilize sequential cyclic translations to learn robust joint representations that may not require all modalities as input at inference time. Tang et al. [15] further explore pairwise bidirectional modality translations. Moreover, Zhao et al. [14] employ data augmentation and cyclic translations based on cascaded residual autoencoder [50] to tackle the uncertain missing modality issue. However, these modality translation-based methods are developed to handle the entire loss of one or more modalities which is less likely to happen in real-world scenarios. Besides, most of them only accept aligned sequences as inputs. Recently, Yuan et al. [16] propose to reconstruct low-level missing features to implicitly force the model to learn semantic features from incomplete multimodal sequences. Lian et al. [51] also introduce low-level feature reconstruction for incomplete multimodal learning in conversations. Compared with these two methods, our proposed method additionally utilizes explicit high-level feature attraction to improve model robustness. Moreover, our experiments demonstrate that explicit high-level feature attraction is superior to implicit low-level feature reconstruction and they work best when combined with each other.

III. Method

In this section, we describe the proposed Efficient Multimodal Transformer with Dual-Level Feature Restoration (EMT-DLFR) for robust MSA, with its whole pipeline in the incomplete modality setting shown in Fig. 1. In this setting, we use the ground-truth complete view to achieve robust representation learning from incomplete multimodal sequences. Specifically, for both views, we first utilize modality-specific encoders to obtain utterance-level (i.e., global) and element-level (i.e., local) intra-modal features from different modality inputs. Then the

Efficient Multimodal Transformer (EMT) is employed to effectively and efficiently capture useful cross-modal interactions between the global multimodal context and local unimodal features. After that, utterance-level intra- and inter-modal features are combined to get the final sentiment intensity prediction. To promote model robustness to random modality feature missing, the Dual-Level Feature Restoration (DLFR) conducts high-level feature attraction and low-level feature reconstruction simultaneously. The former explicitly attracts high-level intra- and inter-modal representations of two views in the latent space, while the latter reconstructs low-level modality inputs from complete view to implicitly urge the model to learn semantic information. In the complete modality setting, the problem degenerates into a simple case. Thus, we do not perform DLFR and only the original prediction loss is used in this setting.

In the following parts, we begin by giving a problem definition. Then we elaborate on the four main modules in EMT-DLFR: unimodal feature encoder (Section III-B), EMT (Section III-C), prediction module (Section III-D), and DLFR (Section III-E). Finally, we present the overall loss function for model training in two modality settings (Section III-F).

A. Problem Definition

The task of MSA is to predict the sentiment intensity of the speaker in the video. Typically, three modalities are involved in MSA, i.e., text or language (l), audio (a), and vision (v). We denote the complete input feature sequences as $\mathring{X}_m \in \mathbb{R}^{T_m \times f_m}$, where T_m is sequence length and f_m is the feature dimension of modality $m \in \{l, a, v\}$. Specifically, $\mathring{X}_a$ and $\mathring{X}_v$ are shallow features extracted by open-source tools, while $\mathring{X}_l$ is the raw text tokens output by the BERT [52] tokenizer. To mimic random modality feature missing in real-world scenarios, we randomly mask the complete sequence $\mathring{X}_m$ to obtain the incomplete sequence $\tilde{X}_m = F(\mathring{X}_m, g_m) \in \mathbb{R}^{T_m \times f_m}$, where $F(\cdot)$ is the mask function, $g_m \in \{0, 1\}^{T_m}$ is a random temporal mask which indicates the positions to mask. For the audio and video modality, the mask function replaces the original feature vector in the masked position with a zero vector. For the text modality, it replaces the original token with the [UNK] token in BERT vocabulary [16]. Our goal is to develop a robust model which can efficiently integrate all the available multimodal information to make an accurate prediction of sentiment intensity score $y \in \mathbb{R}$ in both complete and incomplete modality settings. Note that, when we do not use the diacritical mark $\mathring{}$ or $\tilde{}$ (e.g., X), it indicates that both views can be applied.

B. Unimodal Feature Encoder

Following previous works [16], [29], [32], we use the powerful BERT [52] model to encode raw text tokens into contextual word embeddings. For the audio and video modality, we opt for simplicity and employ the commonly used Long Short-Term Memory (LSTM) recurrent neural network [53] to capture the temporal dependencies in the feature sequence, i.e.,

$$\begin{aligned} H_l &= \text{BERT}(X_l) \\ H_a &= \text{LSTM}(X_a) \\ H_v &= \text{LSTM}(X_v) \end{aligned} \tag{1}$$

where $H_m \in \mathbb{R}^{T_m \times d_m}, m \in \{l, a, v\}$. Since the [CLS] token in BERT aggregates the information from all tokens, we use its embedding as the utterance-level representation of the text modality $h_l \in \mathbb{R}^{d_l}$. For $h_a \in \mathbb{R}^{d_a}$ and $h_v \in \mathbb{R}^{d_v}$, we simply use the feature of the last time step in H_a and H_v. Finally, we use several linear layers to project H_l, H_a, H_v, h_l, h_a, and h_v to the same dimension d to facilitate subsequent multimodal fusion.

C. Efficient Multimodal Transformer

In this part, we first give a preliminary introduction. Then we introduce the basic building block of EMT, named Mutual Promotion Unit (MPU). Based on MPU, we elaborate on three multimodal fusion strategies for modeling cross-modal interactions. The first two are inspired by two predominant models (i.e., MulT [12] and PMR [13]) in MSA. However, both of them suffer from quadratic computational complexity over the involved modalities. The third practically has a linear scaling cost and thus is adopted in EMT by default. Finally, we present hierarchical parameter sharing for EMT to further improve parameter efficiency.

1) Preliminary: Self-attention (SA) is the core component in Transformer. It allows modeling of global dependencies in a sequence via scaled dot-product attention [41]. For the input sequence $H_t \in \mathbb{R}^{T_t \times d}$, we define the Querys as $Q_t = H_t W_Q$, Keys as $K_t = H_t W_K$, Values as $V_t = H_t W_V$, where W_Q and $W_K \in \mathbb{R}^{d \times d_k}$, $W_V \in \mathbb{R}^{d \times d_v}$. Then SA can be formulated as follows:

$$\text{SA}(H_t) = \text{softmax}\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right) V_t \tag{2}$$

Cross-modal attention (CA) involves two modalities. The Querys are from the target modality t, while the Keys and Values are from the source modality s, i.e., $Q_t = H_t W_Q$, $K_s = H_s W_K$, $V_s = H_s W_V$. In this way, CA can provide a latent adaptation from modality s to t:

$$\text{CA}(H_t, H_s) = \text{softmax}\left(\frac{Q_t K_s^T}{\sqrt{d_k}}\right) V_s \tag{3}$$

which is a good way to fuse cross-modal information [12].

Note that, for simplicity, we only present the formulation of single-head attention. In practice, we use multi-head SA or CA (i.e., MHSA or MHCA) to allow the model to attend to information from different feature subspaces [41].

2) Mutual Promotion Unit: The architecture of MPU is shown in Fig. 2. It mainly utilizes symmetric cross-modal attention to explore inherent correlations between elements across two input feature sequences. In this way, MPU enables useful information exchange between two sequences and thus they can *mutually promote* each other. To allow further information integration, self-attention is employed to model temporal dependencies in each feature sequence. Formally, MPU takes two sequences H_m and H_n as inputs and outputs their mutually promoted ones $H_{m \to n}$ and $H_{n \to m}$:

$$\begin{aligned} H_{n \to m}, H_{m \to n} &= \text{MPU}_{m \leftrightarrow n}(H_m, H_n) \\ H_{n \to m} &= \text{MPU}_{n \to m}(H_m, H_n) \\ H_{m \to n} &= \text{MPU}_{m \to n}(H_n, H_m) \end{aligned} \tag{4}$$

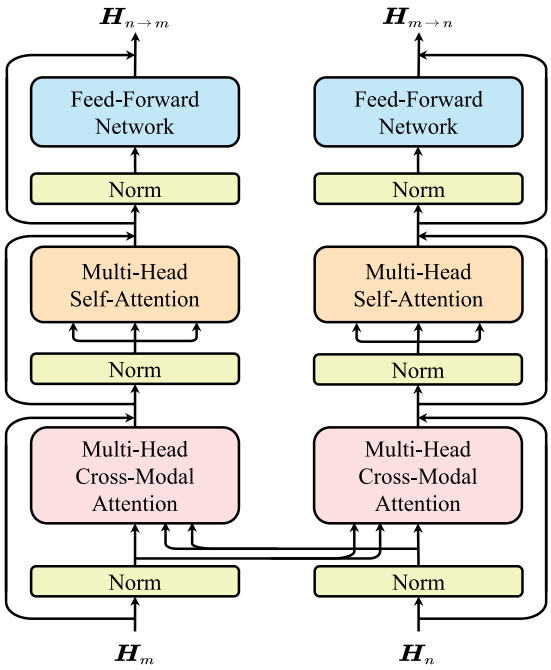


Fig. 2. The architecture of Mutual Promotion Unit (MPU).

TABLE I
COMPUTATIONAL COMPLEXITY ANALYSIS OF THREE MULTIMODAL FUSION STRATEGIES

Fusion Strategy	Space Complexity	Time Complexity
OOLL	$O(M^2)$	$O(M^2T^2)$
OALL	$O(M)$	$O(M^2T^2)$
OAGL	$O(M)$	$O(MT^2)$

M is the number of involved modalities, t is the temporal length of the input feature sequence. For simplicity, we assume aligned multimodal inputs. The detailed time complexity for unaligned inputs can refer to the main text.

where $\to$ and $\leftrightarrow$ indicates the direction of information flow. In specific, the calculation of $MPU_{n\to m}(H_m, H_n)$ is as follows:

$$\begin{aligned} H'_{n\to m} &= \text{MHCA}(\text{LN}(H_m), \text{LN}(H_n)) + H_m, \\ H''_{n\to m} &= \text{MHSA}(\text{LN}(H'_{n\to m})) + H'_{n\to m}, \\ H_{n\to m} &= \text{FFN}(\text{LN}(H''_{n\to m})) + H''_{n\to m} \end{aligned} \tag{5}$$

where LN denotes layer normalization [54], FFN is the position-wise feed-forward network in Transformer. $\text{MPU}_{m\to n}(H_n, H_m)$ can be formulated in a similar way.

Considering that the computational time complexity of $\text{MHCA}(H_m, H_n)$ is $O(T_mT_n)$ and $\text{MHSA}(H_m)$ has the complexity of $O(T_m^2)$, the total time complexity of an MPU is $O(T_mT_n + T_m^2 + T_nT_m + T_n^2) = O((T_m + T_n)^2)$.

3) Multimodal Fusion Strategy: Built upon MPU, we now describe three multimodal fusion strategies in a progressive manner and give a computational complexity analysis for each of them (summarized in Table I). Note that the third fusion strategy (Fig. 1) is adopted in EMT by default and we leave their empirical comparisons in the ablation study (Section V-B).

For simplicity, we only present the information flow from layer i to $i+1$ in an EMT with L layers. We suppose that the input feature sequence is $H_m^{[0]} = H_m + \text{PE}_m \in \mathbb{R}^{T_m\times d}$ and the corresponding utterance-level representation is $h_m^{[0]} = h_m \in \mathbb{R}^d$, where $m \in \{l, a, v\}$ and $\text{PE}_m \in \mathbb{R}^{T_m\times d}$ is the sinusoidal position embedding in the vanilla Transformer.

- *One-to-One Local-Local Fusion:* This fusion strategy shares the same spirit with MulT [12]. The core is to explore *one-to-one local-local* (OOLL) cross-modal interactions across all modalities by applying MPU to each paired modalities, i.e.,

$$\begin{aligned} H_{a\to l}^{[i]}, H_{l\to a}^{[i]} &= \text{MPU}_{l\leftrightarrow a}^{[i]}(H_l^{[i]}, H_a^{[i]}) \\ H_{v\to l}^{[i]}, H_{l\to v}^{[i]} &= \text{MPU}_{l\leftrightarrow v}^{[i]}(H_l^{[i]}, H_v^{[i]}) \\ H_{v\to a}^{[i]}, H_{a\to v}^{[i]} &= \text{MPU}_{a\leftrightarrow v}^{[i]}(H_a^{[i]}, H_v^{[i]}) \end{aligned} \tag{6}$$

  When there are M modalities involved in the fusion process, the number of required MPUs in each layer is $C(M,2) = \frac{1}{2}M(M-1)$, and the time complexity of OOLL is $O(\frac{1}{2}\sum_{m=1}^{M}(\sum_{n\neq m}(T_m + T_n)^2))$. If all modalities are aligned (i.e., $T_m = T, \forall m$), the time complexity degenerates into $O(M^2T^2)$. Thus, OOLL suffers from quadratic space and time complexity over the involved modalities.
- *One-to-All Local-Local Fusion:* Since it is inefficient to fuse multiple modalities in a one-to-one manner, this strategy utilizes the common message introduced in PMR [13] to explore *one-to-all local-local* (OALL) cross-modal interactions in a single MPU. The common message C is initialized by temporal concatenation of unimodal feature sequences, i.e., $C^{[0]} = \text{Concat}(H_l^{[0]}, H_a^{[0]}, H_v^{[0]}) \in \mathbb{R}^{(T_l+T_a+T_v)\times d}$. Formally, we have

$$\begin{aligned} H_l^{[i+1]}, C_{l\to c}^{[i]} &= \text{MPU}_{l\leftrightarrow c}^{[i]}(H_l^{[i]}, C^{[i]}) \\ H_a^{[i+1]}, C_{a\to c}^{[i]} &= \text{MPU}_{a\leftrightarrow c}^{[i]}(H_a^{[i]}, C^{[i]}) \\ H_v^{[i+1]}, C_{v\to c}^{[i]} &= \text{MPU}_{v\leftrightarrow c}^{[i]}(H_v^{[i]}, C^{[i]}) \end{aligned} \tag{7}$$

  Due to the introduction of the common message, OALL only requires M MPUs in each layer. After conducting a similar complexity analysis, we can get the time complexity of OALL: $O(\sum_{m=1}^{M}(T_m + \sum_{n=1}^{M}(T_n))^2)$, which is greater than that of OOLL. In the modality-aligned setting, it degenerates into $O(M^2T^2)$, which indicates that this strategy also has a quadratic scaling cost.
- *One-to-All Global-Local Fusion:* The above analysis indicates that both OOLL and OALL have quadratic computational complexity over the involved modalities. We observe that this issue results from the fact that they both explore local-local cross-modal interactions across all modalities. We believe that it is neither efficient nor necessary to perform multimodal fusion in a local-local fashion. On the one hand, a large amount of redundancy exists in the unimodal feature sequence, especially for the audio and video modality which has high sampling rates. On the other hand, too many local features in one modality (or

the common message) may distract the attention of another modality during their interactions because the latter could not 'see' the global (i.e., summarized) information of the former. Moreover, it could increase the risk of overfitting spurious cross-modal correlation signals. Therefore, we believe that the utterance-level representations from each modality can substitute for the common message in OALL and serves as the *global* multimodal context G to interact with *local* unimodal features, i.e.,

$$\begin{aligned} H_l^{[i+1]}, G_{l\to g}^{[i]} &= \text{MPU}_{l\leftrightarrow g}^{[i]}(H_l^{[i]}, G^{[i]}) \\ H_a^{[i+1]}, G_{a\to g}^{[i]} &= \text{MPU}_{a\leftrightarrow g}^{[i]}(H_a^{[i]}, G^{[i]}) \\ H_v^{[i+1]}, G_{v\to g}^{[i]} &= \text{MPU}_{v\leftrightarrow g}^{[i]}(H_v^{[i]}, G^{[i]}) \end{aligned} \tag{8}$$

where $G^{[0]} = \text{Concat}(h_l^{[0]}, h_a^{[0]}, h_v^{[0]}) \in \mathbb{R}^{3\times d}$. In the above way, this strategy can capture *one-to-all global-local* (OAGL) cross-modal interactions in a single MPU. By stacking multiple layers, the global multimodal context and local unimodal features can mutually promote each other and refine themselves progressively. Same as OALL, OAGL requires M MPUs in each layer. However, thanks to the global multimodal context, the time complexity of OAGL decreases to $O(\sum_{m=1}^{M}(T_m + M)^2) \approx O(\sum_{m=1}^{M} T_m^2)$ (practically, we have $M \ll T_m$), and it degenerates into $O(MT^2)$ in the modality-aligned setting. Thus, the default OAGL fusion strategy in EMT not only has linear space complexity but also enjoys linear computations over the involved modalities.

Finally, we briefly introduce the pooling layer (Fig. 1) for aggregating promoted information from different modalities to facilitate subsequent fusion. Specifically, we utilize an attention-based pooling layer to implement it. Taking OAGL as an example, we have three promoted global multimodal contexts, i.e., $G_{l\to g}^{[i]}$, $G_{a\to g}^{[i]}$, and $G_{v\to g}^{[i]}$. The new global multimodal context can be obtained as follows:

$$G^{[i+1]} = \text{softmax}(v^T \tanh(W^T G_g^T + b)) G_g \tag{9}$$

where $G_g = \text{Concat}(G_{l\to g}^{[i]}, G_{a\to g}^{[i]}, G_{v\to g}^{[i]})$, v, W, and b are learnable parameters. We also develop other two pooling methods, including average pooling and MLP-based pooling. However, they have inferior performance (Section V-B).

4) Hierarchical Parameter Sharing: Inspired by [19], [55], we propose hierarchical parameter sharing to further improve model parameter efficiency. Concretely, there are three levels to share parameters in EMT, including MPU-level sharing, modality-level sharing, and layer-level sharing. The first level shares the parameters of two directional sub-MPUs in an MPU, i.e., $\text{MPU}_{n\to m}$ and $\text{MPU}_{m\to n}$ in (4). The second level shares MPUs across modalities, i.e., $\text{MPU}_{m\leftrightarrow g}^{[i]}$ ($m \in \{l, a, v\}$) in (8). And the last one shares parameters across layers. We show that the hierarchical parameter sharing strategy not only improves parameter efficiency but also can make the optimization of EMT easier in Section V-B.

D. Prediction Module

After multimodal fusion, we flatten the temporal dimension of the final global multimodal context (i.e., $G^{[L]}$) in EMT to get utterance-level inter-modal representation $g \in \mathbb{R}^{3d}$. We then concatenate g and utterance-level intra-modal representations $\{h_m\}$ ($m \in \{l, a, v\}$) and finally pass them through a multi-layer perceptron (MLP) to get the sentiment prediction y'. Following previous works [16], [29], we employ L1 loss as the prediction loss, i.e.,

$$\mathcal{L}_{\text{task}} = |y - y'| \tag{10}$$

E. Dual-Level Feature Restoration

To improve model robustness in the incomplete modality setting, we utilize the complete view to perform the dual-level feature restoration (DLFR), including *implicit* low-level feature reconstruction and *explicit* high-level feature attraction. Note that, the complete view is only used to guide representation learning during training, and it will not be available at the test stage.

1) Low-Level Feature Reconstruction: Reconstruction-based training objective (e.g., masked autoencoding) has achieved great success for representation learning in the fields of computer vision [56], [57], [58], natural language processing [52], [59], and speech signal processing [60], [61]. Motivated by this, we leverage low-level feature restoration (LLFR), i.e., low-level feature reconstruction, to *implicitly* encourage the model to learn semantic representations from incomplete multimodal inputs.

After multimodal fusion via EMT, the incomplete sequence $\tilde{X}_m$ is mapped to a latent sequence $\tilde{Z}_m = \tilde{H}_m^{[L]}$ that has sufficient awareness of both global inter-modal and local intra-modal information. Thus, we pass $\tilde{Z}_m$ through a simple MLP-based decoder $r(\cdot)$ to reconstruct the complete sequence $\mathring{X}_m$. Note that, since reconstructing raw text tokens will waste a large amount of model capacity,[2] we use the output embedding of the BERT model $\mathring{H}_l$ (instead of raw text token sequence $\mathring{X}_l$) as the reconstruction target of the text modality. Following [16], we employ the smooth L1 loss to evaluate the reconstruction quality, i.e.,

$$\begin{aligned} \mathcal{L}_{\text{recon}}^{l} &= \text{smooth}_{\text{L1}}((\mathring{H}_l - r(\tilde{Z}_l)) \cdot (1 - g_l)) \\ \mathcal{L}_{\text{recon}}^{a} &= \text{smooth}_{\text{L1}}((\mathring{X}_a - r(\tilde{Z}_a)) \cdot (1 - g_a)) \\ \mathcal{L}_{\text{recon}}^{v} &= \text{smooth}_{\text{L1}}((\mathring{X}_v - r(\tilde{Z}_v)) \cdot (1 - g_v)) \\ \mathcal{L}_{\text{recon}} &= \sum_{m\in\{l,a,v\}} \mathcal{L}_{\text{recon}}^{m} \end{aligned} \tag{11}$$

where,

$$\text{smooth}_{\text{L1}}(x) = \begin{cases} 0.5x^2 & \text{if } x < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{12}$$

[2]Reconstructing raw text tokens from text representations will need a very big projection matrix $W \in \mathbb{R}^{30522\times 768}$ with about 23.4 M parameters (30522 is the vocabulary size of BERT, 768 is the dimension of text representations).

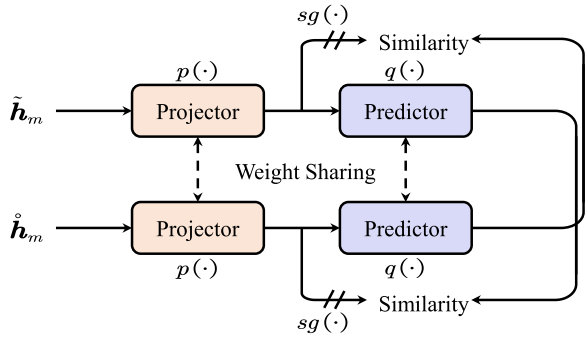


Fig. 3. The illustration of symmetric SimSiam loss.

and the temporal mask g_m ($m \in \{l, a, v\}$) is used to exclude the loss from unmasked positions.

2) High-Level Feature Attraction: As a handcrafted pretext task, low-level feature reconstruction could fail to encourage the model to learn semantic information, as the model may find a shortcut to solve this task (i.e., only utilize local neighbor information to accomplish this task instead of inferring global semantics) [17], [58], [62]. To this end, we further utilize siamese representation learning to *explicitly* attract high-level representations of complete and incomplete views in the latent space. As shown in Fig. 1, except for $\tilde{X}_m$, we also pass $\mathring{X}_m$ through the model to get utterance-level inter-modal representation $\mathring{g}$ and intra-modal representations $\{\mathring{h}_m\}$ of complete view.

The straightforward way to perform feature attraction is to maximize the cosine similarity of representations from two views. However, this method has the risk of collapsing as it may overwhelm the task loss and the model encodes both complete and incomplete inputs to constant vectors [23], [63]. Therefore, we employ a recently proposed self-supervised representation learning framework, SimSiam [23], to avoid possible collapsing solutions.

As shown in Fig. 3, SimSiam consists of an MLP-based projector $p(\cdot)$ and an MLP-based predictor $q(\cdot)$. The input representations of two views are first processed by the projector. Then the predictor maps the intermediate representation of one view to that of the other view. Besides, a stop-gradient operation $sg(\cdot)$ is used to avoid trivial solutions. Formally, we define the symmetric SimSiam loss of utterance-level intra- and inter-modal representations as follows:

$$\begin{aligned}\mathcal{L}_{\text{sim}}^{m} &= \frac{1}{2}[\mathcal{D}(q(p(\tilde{h}_m)), sg(p(\mathring{h}_m))) \\ &\quad + \mathcal{D}(q(p(\mathring{h}_m)), sg(p(\tilde{h}_m)))] \\ \mathcal{L}_{\text{sim}}^{g} &= \frac{1}{2}[\mathcal{D}(q(p(\tilde{g})), sg(p(\mathring{g}))) + \mathcal{D}(q(p(\mathring{g})), sg(p(\tilde{g})))]\end{aligned} \tag{13}$$

where,

$$\mathcal{D}(x, y) = \frac{-x^T y}{||x||_2 ||y||_2} \tag{14}$$

is the negative cosine similarity function, and $m \in \{l, a, v\}$. To ensure the quality of intra- and inter-modal representations from the complete view, we also send $\{\mathring{h}_m\}$ and $\mathring{g}$ to the prediction module and add the corresponding prediction loss in the attraction loss. Thus, the total feature attraction loss is as follows:

$$\mathcal{L}_{\text{attra}} = \sum_{m \in \{l,a,v\}} \mathcal{L}_{\text{sim}}^{m} + \mathcal{L}_{\text{sim}}^{g} + |y - \mathring{y}'| \tag{15}$$

TABLE II
STATISTICS OF THREE MSA BENCHMARK DATASETS

Dataset	Train	Val	Test	All
CMU-MOSI	552/53/679	92/13/124	379/30/277	2199
CMU-MOSEI	4738/3540/8048	506/433/932	1350/1025/2284	23453
CH-SIMS	742/207/419	248/69/139	248/69/140	2281

The three numbers for each split denote the number of the samples with negative (< 0), neutral (= 0), and positive (> 0) sentiment, respectively.

F. Overall Loss Function

In the incomplete modality setting, the task loss is the prediction loss of incomplete view, i.e., $\mathcal{L}_{\text{task}} = |y - \tilde{y}'|$. Thus, the overall loss function can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{attra}} \tag{16}$$

where λ_1 and $\lambda_2 \in \mathbb{R}$ are the weights that balance the contribution of dual-level feature restoration to $\mathcal{L}$.

In the complete modality setting, it is not necessary to perform dual-level feature restoration. Therefore, the overall loss function in this setting is $\mathcal{L} = \mathcal{L}_{\text{task}} = |y - \mathring{y}'|$.

IV. EXPERIMENTS

A. Datasets

In this paper, we conduct extensive experiments on three commonly used benchmark datasets in MSA. We give a brief introduction to each of them and summarize their basic statistics in Table II.

CMU-MOSI: The CMU-MOSI dataset [24] is a widely used MSA benchmark dataset in English. It is a collection of YouTube monologues in which the speakers share their opinions on a wide range of subjects (such as movies). CMU-MOSI consists of a total of 2,199 utterance-level video segments from 93 videos of 89 distinct speakers. Each video segment is manually annotated with a sentiment intensity score, which is defined from -3 (strongly negative) to 3 (strongly positive).

CMU-MOSEI: The CMU-MOSEI dataset [25] is the next generation of CMU-MOSI. Compared with CMU-MOSI, it has much larger training samples and more variations in speakers and video topics. Specifically, CMU-MOSEI contains 23,453 manually annotated utterance-level video segments from 1,000 distinct speakers and 250 different topics.

CH-SIMS: The CH-SIMS dataset [26] is a Chinese MSA benchmark dataset. Compared with the above two datasets, it has both multimodal and unimodal annotations. Nevertheless, we only use the former in this paper. CH-SIMS consists of 2,281 utterance-level video segments from 60 videos whose types span from movies, TV series, and variety shows. Each video segment

is manually annotated with a sentiment intensity score defined from -1 (strongly negative) to 1 (strongly positive).

B. Evaluation Metrics

Since sentiment intensity prediction is primarily a regression task, the typically adopted evaluation metrics are mean absolute error (MAE) and Pearson correlation coefficient (Corr). Researchers also convert the continuous score into different discrete categories and report classification accuracy. Following previous works [12], [16], [25], [32], we report seven-class accuracy (Acc-7), five-class accuracy (Acc-5), binary accuracy (Acc-2), and F1-score on CMU-MOSI and CMU-MOSEI. Note that, there are two distinct methods for the binary formulation, i.e., negative/non-negative [25], and negative/positive [12]. Thus, we report the Acc-2 and F1-score of each method and use a segmentation marker -/- to differentiate them, where the left score is for negative/non-negative and the right score is for negative/positive. On CH-SIMS, following [26], we report five-class accuracy (Acc-5), three-class accuracy (Acc-3), and binary accuracy (Acc-2). For all metrics but MAE, higher values indicate better performance.

To evaluate the model's overall performance under various missing rates in the incomplete modality setting, we follow [16] to compute the Area Under Indicators Line Chart (AUILC) for each of the above metrics. Suppose that the model performance on a metric under increasing missing rates $\{p_0, p_2, \ldots, p_t\}$ is $\{v_0, v_2, \ldots, v_t\}$, the AUILC of this metric is defined as follows:

$$\text{AUILC} = \sum_{i=0}^{t-1} \frac{1}{2}(v_i + v_{i+1})(p_{i+1} - p_i) \tag{17}$$

To be consistent with [16], we evaluate the model under the missing rates of $\{0.1, 0.2, \ldots, 1.0\}$ on CMU-MOSI and CMU-MOSEI, and under the missing rates of $\{0.1, 0.2, \ldots, 0.5\}$ on the CH-SIMS dataset.

C. Feature Extraction

To make fair comparisons, we use the official unaligned features which are provided by the corresponding benchmark datasets and adopted by top-performing MSA methods.

Text Modality: Transformer-based pre-trained language models have achieved state-of-the-art performances on a wide range of tasks in natural language processing. In agreement with recent works [16], [29], [30], we employ the pre-trained BERT model from the open-source Transformers library [64] to encode raw text. Specifically, we use *bert-base-uncased* model for CMU-MOSI and CMU-MOSEI and *bert-base-chinese* model for CH-SIMS.

Audio Modality: For CMU-MOSI and CMU-MOSEI, COVAREP [65] acoustic analysis framework is utilized to extract the low-level acoustic features, which mainly consists of pitch, glottal source parameters, and 12 Mel-frequency cepstral coefficients (MFCCs). For CH-SIMS, an audio and music analysis Python package, Librosa [66], is used to extract logarithmic fundamental frequency, 12 Constant-Q chromatograms, and 20 MFCCs.

TABLE III
DETAILED CONFIGURATIONS ON THREE MSA BENCHMARK DATASETS

Hyper-parameter	CMU-MOSI	CMU-MOSEI	CH-SIMS
Batch size	32	16	32
Learning rate	1e-3	1e-4	1e-3
Learning rate of BERT	5e-5	2e-5	2e-5
Optimizer	Adam	Adam	Adam
Early stop (# epochs)	8	8	8
Gradient accumulation (# batches)	4	4	4
Hidden unit size d in EMT	128	128	32
Stacked layers L in EMT	3	2	4
# attention heads	4	4	4
Expansion factor of FFN	4	4	4
Embedding dropout	0.0	0.0	0.0
Attention dropout	0.3	0.0	0.0
FFN dropout	0.1	0.0	0.0
Loss weight λ_1	1.0	1.0	0.5
Loss weight λ_2	1.0	1.0	0.5

Vision Modality: For CMU-MOSI and CMU-MOSEI, Facet[3] is employed to extract 35 facial action units, which record facial muscle movements related to emotions. For CH-SIMS, OpenFace 2.0 [67] facial behavior analysis toolkit is used to extract 17 facial action units, 68 facial landmarks, and several features related to the head and eyes.

D. Implementation Details

We implement the proposed model using the PyTorch [68] framework. To train the model, we utilize an Adam [69] optimizer and adopt an early-stopping strategy with the patience of 8 epochs. For the hyper-parameters tuning, we perform a random search. The detailed configurations on CMU-MOSI, CMU-MOSEI, and CH-SIMS are summarized in Table III. In the complete modality setting, aligned with [29], [30], we run the model five times and report average results for each evaluation metric. While in the incomplete modality setting, we follow [16] and run the model three times to ensure a fair comparison.

E. Baselines

To comprehensively evaluate the performance of our proposed method in incomplete and complete settings, we consider both general MSA approaches and those which are specially designed to deal with random modality feature missing as our baselines.

TFN: Tensor Fusion Network (TFN) [33] introduces a three-fold Cartesian product-based tensor fusion layer to explicitly model intra-modality and inter-modality dynamics in an end-to-end manner.

LMF: Low-rank Multimodal Fusion (LMF) [34] leverages modality-specific low-rank factors to compute tensor-based multimodal representations, which makes tensor fusion more efficient.

MulT: Multimodal Transformer (MulT) [12] utilizes directional pairwise cross-modal attention to capture inter-modal correlations in unaligned multimodal sequences.

MISA: This method [32] factorizes modalities into Modality-Invariant and -Specific Representations (MISA) using a combination of specially designed losses and then performs multimodal fusion on these representations.

[3]https://imotions.com/platform/

TABLE IV
PERFORMANCE COMPARISON ON CMU-MOSI AND CMU-MOSEI IN THE COMPLETE MODALITY SETTING

	CMU-MOSI						CMU-MOSEI					
Models	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-5 (↑)	Acc-2 (↑)	F1 (↑)	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-5 (↑)	Acc-2 (↑)	F1 (↑)
TFN†	0.901	0.698	34.9	-	-/80.8	-/80.7	0.593	0.700	50.2	-	-/82.5	-/82.1
LMF†	0.917	0.695	33.2	-	-/82.5	-/82.4	0.623	0.677	48.0	-	-/82.0	-/82.1
MulT†	0.861	0.711	-	-	81.5/84.1	80.6/83.9	0.580	0.703	-	-	-/82.5	-/82.3
MISA†	0.804	0.764	-	-	80.8/82.1	80.8/82.0	0.568	0.724	-	-	82.6/84.2	82.7/84.0
Self-MM†	0.712	0.795	45.8	-	82.5/84.8	82.7/84.9	0.529	0.767	53.5	-	82.7/85.0	83.0/84.9
MMIM†	0.700	0.800	46.7	-	84.1/86.1	84.0/86.0	0.526	0.772	54.2	-	82.2/86.0	82.7/86.0
AMML‡	0.723	0.792	46.3	-	-/84.9	-/84.8	0.614	0.776	52.4	-	-/85.3	-/85.2
TFR-Net◇	0.754	0.783	-	54.7	-/84.1	-/-	-	-	-	-	-/-	-/-
MulT	0.846	0.725	40.4	46.7	81.7/83.4	81.9/83.5	0.564	0.731	52.6	54.1	80.5/83.5	80.9/83.6
Self-MM	0.717	0.793	46.4	52.8	82.9/84.6	82.8/84.6	0.533	0.766	53.6	55.4	82.4/85.0	82.8/85.0
MMIM	0.712	0.790	46.9	53.0	**83.3**/**85.3**	**83.4**/**85.4**	0.536	0.764	53.2	55.0	82.5/85.0	82.4/85.1
TFR-Net	0.721	0.789	46.1	53.2	82.7/84.0	82.7/84.0	0.551	0.756	52.3	54.3	81.8/83.5	81.6/83.8
EMT	**0.705**	**0.798**	**47.4**	**54.1**	**83.3**/85.0	83.2/85.0	**0.527**	**0.774**	**54.5**	**56.3**	**83.4**/**86.0**	**83.7**/**86.0**

†: results from [30]. ‡: results from [70]. ◇: results from [16]. All other results are reproduced using publicly available source codes and original hyper-parameters under the same setting. We run each model five times and report average results. For all metrics but MAE, higher values indicate better performance.

Self-MM: Self-Supervised Multi-task Multimodal (Self-MM) sentiment analysis network [29] designs a unimodal label generation module based on self-supervised learning to explore unimodal supervision.

MMIM: MultiModal InfoMax (MMIM) [30] proposes a hierarchical mutual information maximization framework to guide the model to learn shared representations from all modalities.

AMML: Adaptive Multimodal Meta-Learning (AMML) [70] introduces a meta-learning-based method to learn better unimodal representations and then adapt them for subsequent multimodal fusion.

TFR-Net: Transformer-based Feature Reconstruction Network (TFR-Net) [16] employs intra- and inter-modal attention and a feature reconstruction module to deal with random modality feature missing in unaligned multimodal sequences.

TABLE V
PERFORMANCE COMPARISON ON CH-SIMS IN THE COMPLETE MODALITY SETTING

Models	MAE (↓)	Corr (↑)	Acc-5 (↑)	Acc-3 (↑)	Acc-2 (↑)	F1 (↑)
TFN†	0.437	0.582	-	-	77.1	76.9
LMF†	0.438	0.578	-	-	77.4	77.4
MulT†	0.453	0.564	-	-	78.6	79.7
MISA†	0.447	0.563	-	-	76.5	76.6
Self-MM†	0.425	0.595	-	-	80.0	80.4
MulT	0.442	0.581	40.0	65.7	78.2	78.5
Self-MM	0.411	0.601	43.1	66.1	78.6	78.6
MMIM	0.422	0.597	42.0	65.5	78.3	78.2
TFR-Net	0.437	0.583	41.2	64.2	78.0	78.1
EMT	**0.396**	**0.623**	**43.5**	**67.4**	**80.1**	**80.1**

†: results from [71]. All other results are reproduced using publicly available source codes and original hyper-parameters under the same setting. We run each model five times and report average results. For all metrics but MAE, higher values indicate better performance.

V. RESULTS

A. Comparison to State-of-the-Art

1) Complete Modality Setting: First, we compare the proposed EMT with top-performing MSA methods in the complete modality setting. Table IV shows the results on CMU-MOSI and CMU-MOSEI. Since the baseline papers do not have a common evaluation setting and only report results on partial metrics, we present both the results from original papers and those reproduced by ourselves under the same setting (they are generally comparable). For fair and full comparisons, we mainly compare ours with the reproduced one. From Table IV, we first observe that our EMT shows superior performance over previous state-of-the-art Transformer-based methods (i.e., TFR-Net, and MulT). Specifically, it surpasses the best-performing TFR-Net by 0.016 MAE on CMU-MOSI and 2.2% Acc-7 on CMU-MOSEI, and outperforms MulT by 0.141 MAE on CMU-MOSI and 0.043 Corr on CMU-MOSEI. We can attribute these encouraging results to the effective yet efficient global-local cross-modal interaction modeling in EMT, because both two baseline methods focus on capturing pairwise local-local cross-modal interactions, which not only suffer from a large amount of redundancy but also increase the risk of overfitting. In addition to Transformer-based methods, EMT also surpasses several recent state-of-the-art non-Transformer-based methods (e.g., AMML, MMIM, and Self-MM), setting new records on most metrics. Finally, we present the results on the Chinese MSA dataset CH-SIMS in Table V. It can be seen that EMT achieves the best performance on all metrics. For instance, it outperforms the second performer Self-MM by 0.015 MAE, 0.022 Corr, and 1.5% Acc-2. In summary, the above results demonstrate the effectiveness of EMT in the complete modality setting.

2) Incomplete Modality Setting: Next, we concentrate on evaluating the robustness of our EMT-DLFR under various missing rates. Note that, to generate incomplete views, we apply the same missing rate to three modalities during both the training and test stages. Besides, we independently generate the random temporal mask for each modality. The performance comparison on CMU-MOSI, CMU-MOSEI, and CH-SIMS are shown in Fig. 4, 5, and 6, respectively. For simplicity, we only show four representative metrics (i.e., MAE, Corr, Acc-7/Acc-5, and Acc-2) for each dataset. Note that the Acc-2 on CMU-MOSI and CMU-MOSEI means the binary accuracy calculated on negative/positive samples. From the three figures, we have the following observations: (1) In general, with the increase in missing rates, the performances of all methods decrease

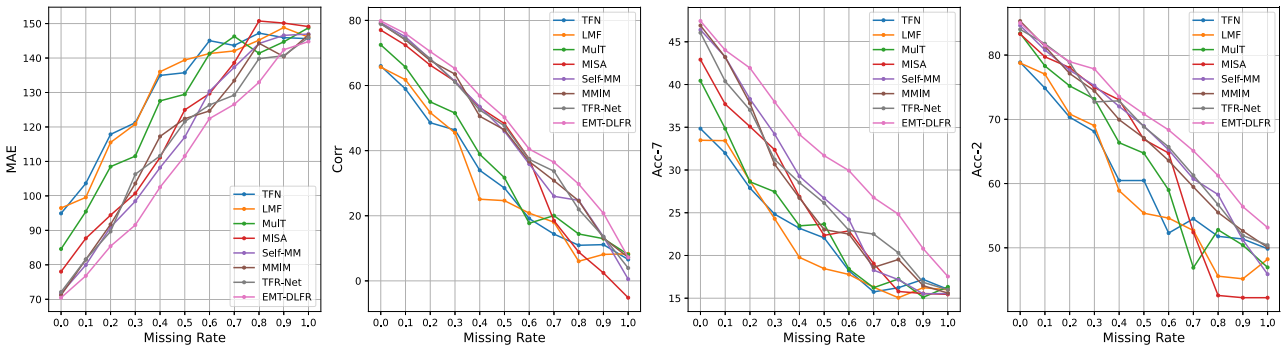


Fig. 4. Performance comparison under various missing rates on CMU-MOSI.

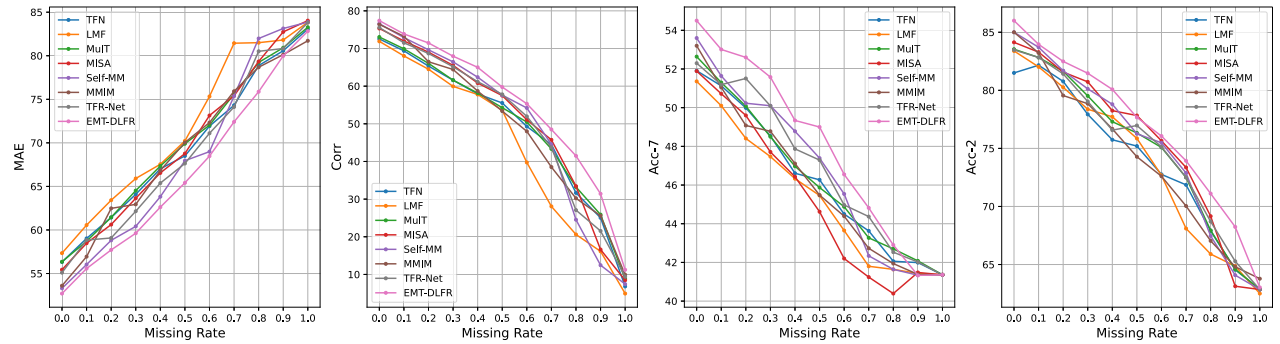


Fig. 5. Performance comparison under various missing rates on CMU-MOSEI.

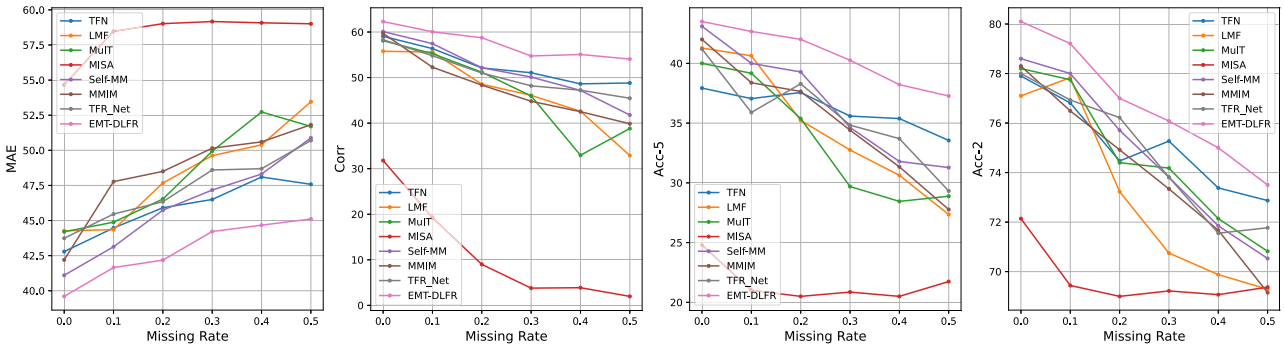


Fig. 6. Performance comparison under various missing rates on CH-SIMS.

quasi-linearly. It indicates that random modality feature missing in unaligned multimodal sequences will moderately degrade the model's performance and this issue needs to be carefully addressed in the real-world deployment of MSA models. (2) EMT-DLFR outperforms all other compared methods in most cases on three datasets, especially when the missing rate is relatively high, thus clearly illustrating the robustness of our proposed method in the incomplete modality setting.

To quantitatively evaluate the overall performance of different methods, we follow [16] to calculate the AUILC of each evaluation metric, as stated in Section IV-B. The results on CMU-MOSI and CMU-MOSEI are shown in Table VI. Similar to the complete case, we report both original and reproduced results, and we mainly compare ours with the latter. On CMU-MOSI, we can find that the state-of-the-art TFR-Net which is specialized to deal with random modality feature missing outperforms other general MSA methods in most cases. This indicates that low-level feature reconstruction adopted by TFR-Net can indeed force the model to learn certain semantic information from incomplete multimodal sequences. Nevertheless, our EMT-DLFR which utilizes both low-level feature reconstruction and high-level feature attraction improves both TFR-Net and other methods by a large margin on almost all metrics. To be specific, it outperforms the second performer by 0.050 MAE,

TABLE VI
OVERALL PERFORMANCE COMPARISON ON CMU-MOSI AND CMU-MOSEI IN THE INCOMPLETE MODALITY SETTING

	CMU-MOSI						CMU-MOSEI					
Models	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-5 (↑)	Acc-2 (↑)	F1 (↑)	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-5 (↑)	Acc-2 (↑)	F1 (↑)
TFN$^\diamond$	1.327	0.300	-	23.3	-/60.4	-/-	-	-	-	-	-/-	-/-
MulT$^\diamond$	1.288	0.334	-	24.4	-/61.8	-/-	-	-	-	-	-/-	-/-
MISA$^\diamond$	1.209	0.403	-	27.1	-/63.2	-/-	-	-	-	-	-/-	-/-
TFR-Net$^\diamond$	1.155	0.467	-	30.4	-/69.0	-/-	-	-	-	-	-/-	-/-
TFN	1.316	0.308	22.3	23.7	61.0/60.9	59.7/59.7	0.695	0.500	46.1	46.6	75.2/74.1	73.4/71.5
LMF	1.310	0.299	21.5	22.7	59.7/59.3	56.4/56.1	0.718	0.447	45.3	45.7	72.2/73.9	69.5/69.4
MulT	1.263	0.348	23.1	24.6	63.1/63.2	60.7/61.0	0.700	0.504	46.3	46.8	74.4/75.1	72.9/72.6
MISA	1.202	0.405	25.7	27.4	63.9/63.7	59.0/58.8	0.698	0.514	45.1	45.7	75.2/75.7	74.4/74.0
Self-MM	1.162	0.444	27.8	30.3	66.9/67.5	65.4/66.2	0.685	0.507	46.7	47.3	75.1/75.4	73.7/72.9
MMIM	1.168	0.450	27.0	29.4	66.8/66.9	64.6/65.8	0.694	0.502	45.9	46.4	74.9/72.4	72.4/69.3
TFR-Net	1.156	0.452	27.7	30.5	67.6/67.8	65.7/66.1	0.689	0.511	46.9	47.3	74.7/74.2	73.5/73.4
EMT-DLFR	**1.106**	**0.486**	**32.5**	**35.6**	**69.6/70.3**	**69.6/70.3**	**0.665**	**0.546**	**47.9**	**48.8**	**76.4/76.9**	**75.2/75.9**

The reported result is the AUILC of each evaluation metric, which is calculated under the missing rates of $\{0.1, 0.2, \ldots, 1.0\}$. $\diamond$: results from [16]. All other results are reproduced using publicly available source codes and original hyper-parameters under the same setting. We run each model three times and report average results. For all metrics but MAE, higher values indicate better performance.

TABLE VII
OVERALL PERFORMANCE COMPARISON ON CH-SIMS IN THE INCOMPLETE MODALITY SETTING

Models	MAE (↓)	Corr (↑)	Acc-5 (↑)	Acc-3 (↑)	Acc-2 (↑)	F1 (↑)
TFN$^\diamond$	0.233	0.259	18.1	-	37.3	-
MulT$^\diamond$	0.244	0.227	17.3	-	37.0	-
MISA$^\diamond$	0.294	0.038	10.6	-	34.7	-
TFR-Net$^\diamond$	0.237	0.253	18.0	-	37.7	-
TFN	0.230	0.262	18.1	30.5	37.5	37.5
LMF	0.241	0.237	17.4	30.1	36.5	36.1
MulT	0.242	0.233	16.7	30.0	37.3	36.8
MISA	0.293	0.053	10.6	26.4	34.8	28.9
Self-MM	0.231	0.258	18.3	30.5	37.4	37.5
MMIM	0.244	0.238	17.7	29.8	37.0	36.2
TFR-Net	0.236	0.253	17.8	30.0	37.3	37.2
EMT-DLFR	**0.215**	**0.287**	**20.4**	**31.9**	**38.4**	**38.5**

The reported results are the AUILC of each evaluation metric, which are calculated under the missing rates of $\{0.1, 0.2, \ldots, 0.5\}$. $\diamond$: results from [16]. All other results are reproduced using publicly available source codes and original hyper-parameters under the same setting. We run each model three times and report average results. For all metrics but MAE, higher values indicate better performance.

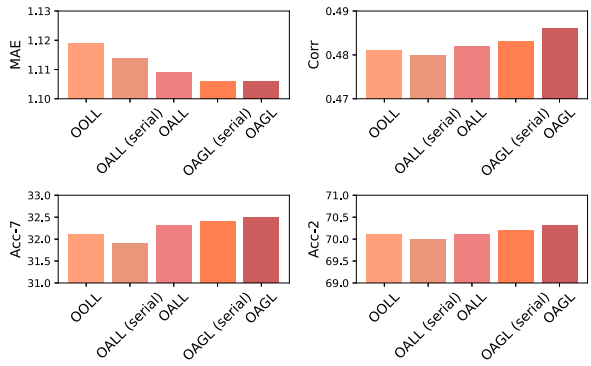


Fig. 7. Ablation study of multimodal fusion strategy on CMU-MOSI in the incomplete modality setting.

0.034 Corr, 4.7% Acc-7, 5.1% Acc-5, and 3.9%/4.1% F1, which amply demonstrates the superiority of the proposed method. On CMU-MOSEI, we observe that the difference between different methods is smaller than that on CMU-MOSI (also shown in Fig. 5), which could be partly ascribed to much larger training samples in this dataset. Our EMT-DLFR still achieves 0.020 and 0.032 higher performance on MAE and Corr than the second performer. We further present the results on CH-SIMS in Table VII. The proposed method also achieves superior performance, outperforming the previous state-of-art by 0.015 MAE and 0.025 Corr. To summarize, the above results verify the robustness of our proposed method in the incomplete modality setting.

B. Ablation Study

To further investigate the influence of each component in our method, we conduct comprehensive ablation experiments on CMU-MOSI in the incomplete modality setting.

1) Multimodal Fusion Strategy: We first validate the effectiveness and efficiency of the proposed multimodal fusion strategy. We compare the default OAGL fusion strategy adopted by EMT with two baseline fusion strategies (i.e., OOLL and OALL) inspired by previous methods as stated in Section III-C. We show the performance of three fusion strategies in Fig. 7. We can find that OAGL outperforms OOLL and OALL, which verifies the feasibility and effectiveness of utilizing utterance-level representations from each modality as the global multimodal context to interact with local unimodal features. We think that the superiority of OAGL mainly comes from the simplification of cross-modal interaction modeling. In comparison to the local-local version in OOLL and OALL, it not only minimizes redundant information but also encourages the model to concentrate exclusively on prominent cross-modal correlation patterns, thus making it less likely to overfit the spurious ones. Additionally, inspired by [13], for OALL and OAGL, we develop their serial variants by replacing H_m in the last row of (4) with $H_{n\to m}$ (i.e., the promoted H_m) in the second row. From Fig. 7, we observe that the serial variant performs a bit worse than the default one for both OALL and OAGL, which justifies the parallel design choice of two fusion strategies. We believe this might be due to the loss of original unimodal information brought by the too-fast feature update in the serial implementation.

To further illustrate the efficiency of OAGL over OOLL and OALL, we give a practical complexity comparison of them.

TABLE VIII
THE COMPLEXITY COMPARISON OF DIFFERENT MULTIMODAL FUSION STRATEGIES ON CMU-MOSI IN THE INCOMPLETE MODALITY SETTING

Fusion Strategy	MACs (G)	#Params (M)	Training Time (s)	GPU Memory (GB)
MulT	-	111.0	17.5	17.8
TFR-Net	-	124.3	24.8	16.9
OOLL	3.1	**110.5**	17.1	17.8
OALL	8.3	**110.5**	21.2	31.5
OAGL	**1.5**	**110.5**	**15.4**	**10.8**

Since there are three input modalities, these three strategies have the same space complexity (as they all need 3 MPUs in each fusion layer). Thus, we mainly evaluate the time complexity in this paper. Nevertheless, it is worth noting that OAGL has linear scalability of MPUs when more modalities are involved in the fusion process. In Table VIII, we compare three fusion strategies in terms of the amount of computation measured in multiply–accumulate operations (MACs),[4] the number of parameters, one epoch training time, and GPU memory usage during model training. Note that the MACs are only calculated on the fusion module, excluding the other parts (e.g., unimodal feature encoders) of the model. For the number of parameters, we do not count the networks for feature reconstruction and attraction because they will not be used during inference. We run each experiment five times using a batch size of 32 on a single Tesla V100 GPU (32 GB) and report the mean value for each metric. In addition to OOLL and OALL, we also provide metrics of two state-of-the-art multimodal Transformers as references, including MulT and TFR-Net.

From Table VIII, we can observe that OAGL has the least MACs among the three fusion strategies (5× less than OALL and 2× less than OOLL). This is expected since OAGL enjoys linear computational complexity over the involved modalities in practice while both OOLL and OALL have a quadratic scaling cost. Compared with OOLL, OALL which utilizes the common message (i.e., local multimodal context) to explore multi-way cross-modal interactions achieves better performance (as shown in Fig. 7) but at the expense of a much larger amount of computation (especially the GPU memory usage). In contrast, thanks to the introduction of the global multimodal context, our OAGL not only has the best performance but also has the least computation. We also notice that the improvement in training speed is less significant than that of MACs. This is owing to the dominant role of the text modality feature encoder (i.e., BERT) in the overall computational overhead. Finally, the state-of-the-art TFR-Net and MulT, which utilize similar strategies to OOLL for multimodal fusion, also have much larger computational costs and more parameters than our OAGL. To summarize, the above quantitative results demonstrate the efficiency of OAGL over traditional local-local methods for cross-modal interaction modeling.

2) Hierarchical Parameter Sharing for EMT: To verify the role of hierarchical parameter sharing for EMT, we evaluate different combinations of three-level (i.e., MPU-level, modality-level, and layer-level) parameter sharing strategies in Table IX. Surprisingly, we find that all methods have comparable performance and the full (i.e., not-shared) model is even slightly inferior to several parameter-sharing variants. Moreover, the all-shared model achieves decent results only with negligible performance degradation on MAE when compared to the full model. These results indicate that EMT is more difficult to train when there are too many parameters, and parameter sharing can act as a regularization constraint to alleviate this problem to some extent. Another benefit of parameter sharing is the big improvement in parameter efficiency. Note that, compared with the not-shared case, sharing all parameters in EMT leads to a reduction of 6 M parameters. Since the text BERT encoder has 109.5 M parameters, there are only less than 1 M (actually, 0.5 M) parameters in EMT for the all-shared model.

[4]https://github.com/Lyken17/pytorch-OpCounter

TABLE IX
ABLATION STUDY OF HIERARCHICAL PARAMETER SHARING IN EMT ON CMU-MOSI IN THE INCOMPLETE MODALITY SETTING

MPU	Modality	Layer	#Params (M)	MAE (↓)	Corr (↑)
×	×	×	116.4	1.105	0.486
✓	×	×	113.4	1.103	0.484
×	✓	×	112.4	1.105	**0.489**
×	×	✓	112.1	1.104	0.487
✓	✓	×	111.4	1.103	0.487
✓	×	✓	111.1	1.106	0.485
×	✓	✓	110.8	**1.101**	0.488
✓	✓	✓	**110.5**	1.106	0.486

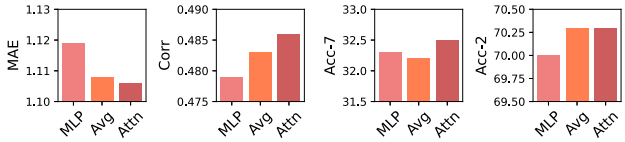


Fig. 8. Ablation study of pooling layer in EMT on CMU-MOSI in the incomplete modality setting.

3) Pooling Layer in EMT: In this part, we ablate the design choice of the pooling layer in EMT. The pooling layer in EMT is used to aggregate promoted information from different modalities. We compare the default attention pooling with two variants, i.e., average pooling and MLP-based pooling. The results are shown in Fig. 8. As a non-parametric method, average pooling is slightly inferior to attention pooling since it can not make aware of the contributions of different modalities. MLP-based pooling directly maps multiple promoted features into a single one using a fully connected layer. Thus, compared with attention-pooling, it has much more parameters. However, MLP-based pooling achieves the worst performance among the three methods. We guess this could be caused by the overfitting problem considering that MSA datasets are relatively small.

4) Dual-Level Feature Restoration: We then investigate the contributions of low-level feature reconstruction and high-level feature attraction to the improvement of model robustness. We compare our full method EMT-DLFR with the following three variants:

- *EMT-LLFR:* only use low-level feature reconstruction by removing $\mathcal{L}_{\text{attra}}$ in (16).

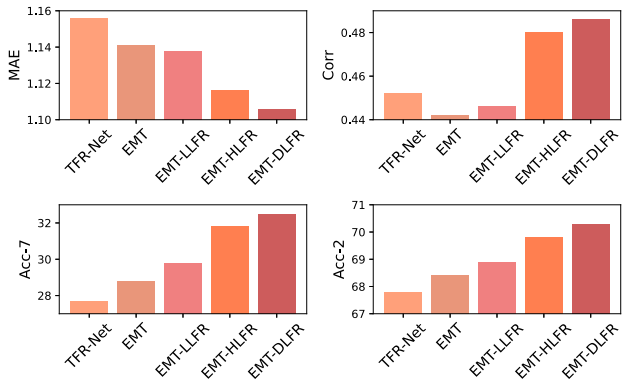


Fig. 9. Ablation study of loss function on CMU-MOSI in the incomplete modality setting.

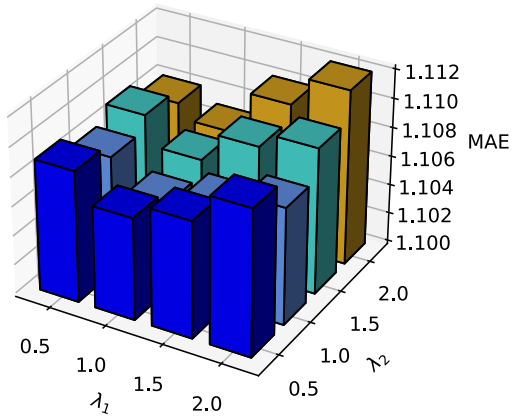


Fig. 10. Ablation study of loss weights on CMU-MOSI in the incomplete modality setting.

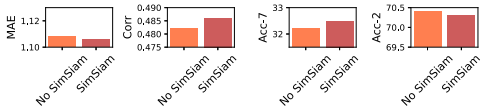


Fig. 11. Ablation study of SimSiam for high-level feature attraction on CMU-MOSI in the incomplete modality setting.

- *EMT-HLFR:* only use high-level feature attraction by removing $\mathcal{L}_{\text{recon}}$ in (16).
- *EMT:* use neither low-level feature reconstruction nor high-level feature attraction by removing both $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{attra}}$ in (16).

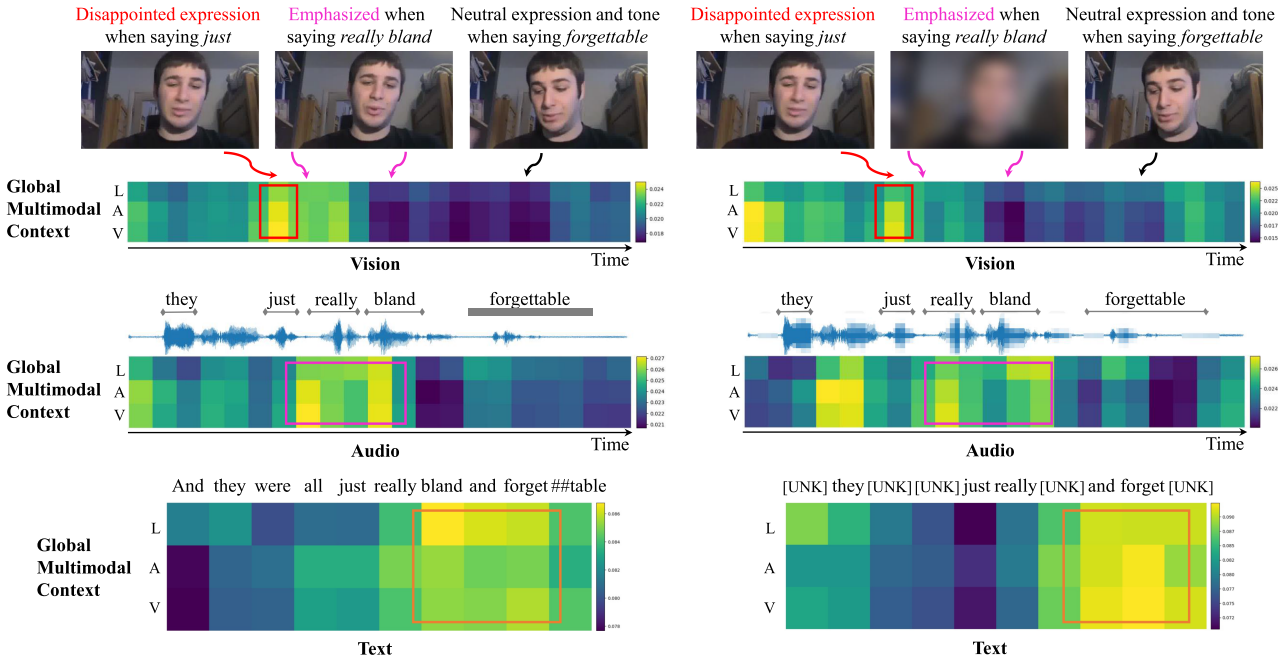
Fig. 9 shows the results of different methods. We can observe that EMT-LLFR and EMT-HLFR outperform vanilla EMT, which indicates both low-level feature reconstruction and high-level feature attraction can improve model robustness to random modality feature missing. Besides, high-level feature attraction is more crucial than low-level feature reconstruction considering that EMT-LLFR significantly underperforms EMT-HLFR by 0.022 MAE, 0.034 Corr, 2.0% Acc-7, and 0.9% Acc-2. This is conceptually intuitive since implicit low-level feature reconstruction might not be sufficient to force the model to infer high-level semantics, as it could find a shortcut (i.e., only using local neighbor information) to accomplish the reconstruction task. In contrast, directly performing explicit attraction of high-level features from the complete and incomplete view in the latent space is more beneficial and effective. Moreover, our full method EMT-DLFR achieves the best performance among all methods, which suggests that implicit low-level feature reconstruction and explicit high-level feature attraction are complementary to each other. Therefore, the combination of these two feature restoration mechanisms can be a unified framework for robust multimodal sentiment analysis. For reference, we also present the result of state-of-the-art TFR-Net in Fig. 9. As expected, both EMT-LLFR and EMT-HLFR have better performance than TFR-Net. The encouraging thing is that even vanilla EMT slightly outperforms TFR-Net on several metrics (e.g., MAE, Acc-7, and Acc-2), which once again demonstrates the superiority of EMT over traditional local-local cross-modal fusion methods.

5) Sensitivity of Loss Weights: After ablating the contribution of two feature restoration methods, we investigate the sensitivity of their corresponding weights in the loss function, i.e., λ_1 for low-level feature reconstruction and λ_2 for high-level feature attraction in (16). We try several values for each weight parameter in the range of 0.5 to 2.0 with a step of 0.5. The results of different combinations of λ_1 and λ_2 are shown in Fig. 10. We mainly have the following observations: 1) All weight combinations have comparable performance, which suggests that our proposed method is insensitive to the values of λ_1 and λ_2. 2) Too large weights may hurt the performance. 3) When $\lambda_1 = 1.0$ and $\lambda_2 = 1.0$, the model achieves the best performance. Thus, we use them as the default loss weights on this dataset.

6) SimSiam for High-Level Feature Attraction: Finally, we analyze the impact of SimSiam in siamese representation learning for explicit high-level feature attraction. The alternative to SimSiam is to simply minimize the negative cosine similarity of utterance-level intra- and inter-modal representations between complete and incomplete views. We present the results of two methods in Fig. 11. We observe that the performance generally decreases when SimSiam is not used in siamese representation learning, which verifies the role of SimSiam in preventing the model from learning collapsed representations by virtue of its special architecture designs.

C. Visualization Analysis

To have a deeper understanding of how our model works when modeling unaligned multimodal sequences in both complete and incomplete modality settings, we empirically investigate the signals EMT captures by visualizing the cross-modal attention weights. Fig. 12 presents the global-local cross-modal attention matrices of a sample selected from the test set of CMU-MOSI. For convenience, we only show the attention of the global multimodal context to each unimodal feature (i.e., $m \rightarrow g$, $m \in \{l, a, v\}$ in (8)). Since the acoustic and visual sequences are too long (42 and 50 respectively), we use average pooling with a stride size of 2 to halve them. We also show the key video



(a) Complete Modality Setting (b) Incomplete Modality Setting

Fig. 12. Visualization of global-local cross-modal attention weights from EMT in the complete and incomplete (under a missing rate of 0.5) modality setting. The sample jUzDDGyPkXU_27 in the test set of CMU-MOSI is selected. For convenience, we only show the attention of the global multimodal context to local unimodal features (i.e., the global multimodal context is the target while each local unimodal feature is the source). High and meaningful attention areas are highlighted by colored rectangles. For both settings, we find that the global multimodal context has learned to pay attention to meaningful signals in each modality (e.g., disappointed facial expression, emphasized tone, and sentiment words).

frames, audio waveform, text tokens from BERT, and alignment information for better interpretation. It should be noted that, due to the segmentation error, the first word *And* was not said by the speaker though it appears in the transcription provided by the dataset.

As shown in Fig. 12(a), we find that our model captures meaningful global-local cross-modal interactions in the complete modality setting. Specifically, the global multimodal context pays its most attention to the visual segments where a disappointed facial expression occurs. For the audio modality, it learns to attend to the emphasized intervals in the acoustic sequence. While for the high-level text modality, we can see that sentiment words (e.g., *bland* and *forgettable*) are received stronger attention from the global multimodal context. These interesting observations qualitatively demonstrate the effectiveness of EMT for cross-modal interaction modeling. Finally, our model predicts a negative sentiment score of -2.1 for this sample, which is very close to the ground-truth label -2.0.

In the incomplete modality setting, we use a missing rate of 0.5 to generate incomplete multimodal sequences. The missing audio and visual features are padded with zeros. The missing text tokens are replaced by the unknown token [UNK] in BERT. In Fig. 12(b), we use blur background and mosaic to indicate the missing vision and audio modality features. It should be noted that they might not coincide with the real-generated temporal masks since they are only used for illustrative purposes. From Fig. 12(b), we observe that, although the attention matrices are more scattered due to the influence of random modality feature missing, our model still attends to those useful cross-modal signals captured in the complete modality setting. Notably, for the text modality, we find that the model pays more attention to the available token *forget* as expected but also gives partial attention to the missing tokens (e.g., *bland* and *##table*). Under such a moderate missing rate, our model prediction remains a negative score of -1.0. These results once again verify the efficacy of the proposed dual-level feature restoration to the improvement of model robustness.

In addition, we also visualize the local-local cross-modal attention weights to empirically investigate the differences between the proposed global-local fusion strategy (i.e., OAGL) and the previous local-local one (e.g., OOLL and OALL). For simplicity, we only show OOLL in Fig. 13 as we have similar observations for OALL. We find that, although OOLL can capture a part of meaningful cross-modal correlation signals as OAGL, its attention matrices are typically low-rank (especially for the upper part in Fig. 13), which demonstrates that a large amount of redundancy exists in local-local cross-modal interactions and it can be greatly reduced via the global-local interactions in OAGL to achieve efficient multimodal fusion. This empirical finding is also consistent with that in the previous study [12]. Moreover, we notice that OOLL attends to more irrelevant cross-modal information than OAGL (such as the relatively higher attention to audio and vision modalities when saying *forgettable*), implying that the local-local fusion strategy could also increase the risk of overfitting the spurious correlations in unaligned multimodal data.

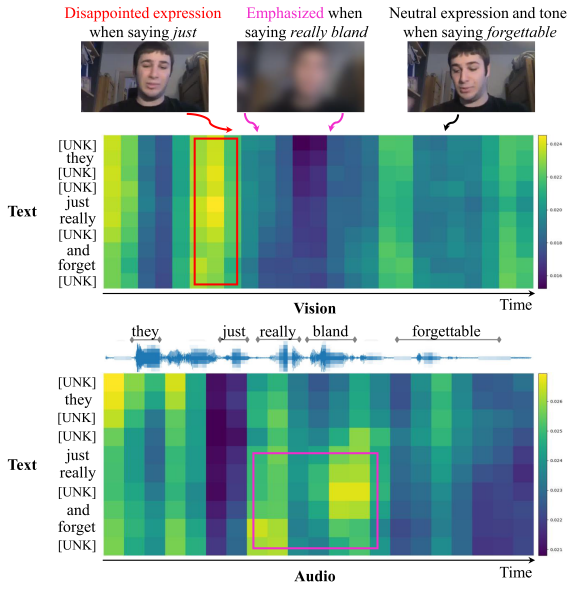


Fig. 13. Visualization of local-local cross-modal attention weights from EMT (OOLL) in the incomplete modality setting (under a missing rate of 0.5). For convenience, we only show the attention of the text modality to audio and vision modalities. High and meaningful attention areas are highlighted by colored rectangles.

VI. Conclusion

In this article, we have presented a generic and unified framework, named Efficient Multimodal Transformer with Dual-Level Feature Restoration (EMT-DLFR), for efficient and robust multimodal sentiment analysis. At the heart of EMT is the introduction of the global multimodal context, which enables effective and efficient exploration of global-local cross-modal interactions. It not only avoids the quadratic scaling cost of previous local-local cross-modal interaction modeling methods but also leads to performance gains. Furthermore, we utilize hierarchical parameter sharing to improve parameter efficiency and ease model training. To cope with random modality feature missing which typically occurs in realistic settings, DLFR employs both implicit low-level feature reconstruction and explicit high-level feature attraction to achieve robust representation learning from incomplete multimodal data. We find that, although the former is more effective than the latter, these two strategies are complementary to each other and thus can be combined to achieve better performance. Finally, extensive experiments on three datasets show that the proposed method achieves state-of-the-art performance in both complete and incomplete modality settings. In addition, empirical visualization analysis also demonstrates that our model can capture interpretable and robust cross-modal correlation signals for reliable sentiment prediction.

In future work, we hope to apply EMT-DLFR to the entire modality missing setting, noisy modality setting, and even more challenging noise-missing-mixed setting. It's also interesting to explore the effectiveness of the proposed method on different multimodal learning tasks and datasets. Besides, since we do not take into consideration the importance of each modality, it would be helpful to explicitly incorporate it into EMT-DLFR to further improve performance. Finally, although our EMT enjoys a linear scaling cost over the involved modalities, it still suffers from quadratic complexity with respect to the input sequence length. Thus, how to further reduce the complexity from $O(MT^2)$ to $O(MT)$ remains an attractive research direction.

References

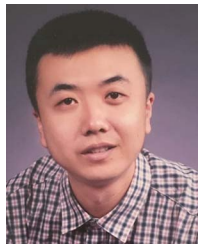
[1] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, pp. 3–14, 2017.
[2] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 108–132, First Quarter 2023.
[3] K. Somandepalli, T. Guha, V. R. Martinez, N. Kumar, H. Adam, and S. Narayanan, "Computational media intelligence: Human-centered machine analysis of media," in *Proc. IEEE*, vol. 109, no. 5, pp. 891–910, May 2021.
[4] L. Stappen, A. Baird, L. Schumann, and S. Bjorn, "The multimodal sentiment analysis in car reviews (MuSe-CaR) dataset: Collection, insights and improvements," *IEEE Trans. Affect. Comput.*, vol. 14, no. 2, pp. 1334–1350, Second Quarter 2023.
[5] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," in *A Practical Guide to Sentiment Analysis*. Berlin, Germany: Springer, 2017, pp. 1–10.
[6] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, 2017.
[7] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 873–883.
[8] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.
[9] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5642–5649.
[10] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. Conf. Assoc. Comput. Linguistics*, 2018, Art. no. 2225.
[11] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6892–6899.
[12] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.
[13] F. Lv, X. Chen, Y. Huang, L. Duan, and G. Lin, "Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2554–2562.
[14] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 2608–2618.
[15] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5301–5311.
[16] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 4400–4407.
[17] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021.
[18] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2019, pp. 3744–3753.
[19] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 4651–4664.

[20] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 14 200–14 213.
[21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2020, pp. 1597–1607.
[23] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15 750–15 758.
[24] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov./Dec. 2016.
[25] A. Zadeh and P. Pu, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[26] W. Yu et al., "CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3718–3727.
[27] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 169–176.
[28] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, "Utterance-level multimodal sentiment analysis," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, 2013, pp. 973–982.
[29] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10 790–10 797.
[30] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9180–9192.
[31] Z. Lian, J. Tao, B. Liu, and J. Huang, "Conversational emotion analysis via attention mechanisms," in *Proc. Int. Speech Commun. Assoc.*, 2019, pp. 1936–1940.
[32] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.
[33] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.
[34] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.
[35] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 481–492.
[36] T. Jin, S. Huang, Y. Li, and Z. Zhang, "Dual low-rank multimodal fusion," in *Proc. Findings Assoc. Comput. Linguistics*, 2020, pp. 377–387.
[37] S. S. Rajagopalan, L.-P. Morency, T. Baltrusaitis, and R. Goecke, "Extending long short-term memory for multi-view structured learning," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 338–353.
[38] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. 19th ACM Int. Conf. Multimodal Interaction*, 2017, pp. 163–171.
[39] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7216–7223.
[40] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.
[41] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[42] A. Zadeh et al., "Factorized multimodal transformer for multimodal sequential learning," 2019, *arXiv: 1911.09826*.
[43] T. Liang, G. Lin, L. Feng, Y. Zhang, and F. Lv, "Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8148–8156.
[44] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross-and self-attention network for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 4275–4279.
[45] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 985–1000, 2021.
[46] V. Rajan, A. Brutti, and A. Cavallaro, "Is cross-attention preferable to self-attention for multi-modal emotion recognition?," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 4693–4697.
[47] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *Proc. Int. Conf. Multimodal Interact.*, 2020, pp. 400–404.
[48] D. Hazarika, Y. Li, B. Cheng, S. Zhao, R. Zimmermann, and S. Poria, "Analyzing modality robustness in multimodal sentiment analysis," 2022, *arXiv:2205.15465*.
[49] P. P. Liang, Z. Liu, Y.-H. H. Tsai, Q. Zhao, R. Salakhutdinov, and L.-P. Morency, "Learning representations from imperfect time series data via tensor rank regularization," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1569–1576.
[50] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1405–1414.
[51] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, "GCNet: Graph completion network for incomplete multimodal learning in conversation," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2023.3234553.
[52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol.*, Minneapolis, Minnesota, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19--1423
[53] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[54] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[55] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 1–17.
[56] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.
[57] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2536–2544.
[58] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16 000–16 009.
[59] A. Radford et al., "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
[60] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 12 449–12 460.
[61] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.
[62] H. Bao, L. Dong, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.
[63] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21 271–21 284.
[64] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process. Syst. Demonstrations*, 2020, pp. 38–45.
[65] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP–a collaborative voice analysis repository for speech technologies," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 960–964.
[66] B. McFee et al., "Librosa: Audio and music signal analysis in python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.
[67] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. IEEE 13th Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.

[68] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[69] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[70] Y. Sun, S. Mai, and H. Hu, "Learning to learn better unimodal representations via adaptive multimodal meta-learning," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2022.3178231.
[71] H. Mao, Z. Yuan, H. Xu, W. Yu, Y. Liu, and K. Gao, "M-SENA: An integrated platform for multimodal sentiment analysis," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics Syst. Demonstrations*, 2022, pp. 204–213.

Licai Sun received the BS degree from Beijing Forestry University, Beijing, China, in 2016, and the MS degree from the University of Chinese Academy of Sciences, Beijing, China, in 2019. He is currently working toward the PhD degree with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China. His current research interests include affective computing, deep learning, and multimodal representation learning.

Zheng Lian received the BS degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2016 and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2021. He is currently an assistant professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include affective computing, deep learning, and multimodal emotion recognition.



Bin Liu (Member, IEEE) received the BS and MS degrees from the Beijing institute of technology (BIT), Beijing, China, in 2007 and 2009 respectively and the PhD degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015. He is currently an associate professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests include affective computing and audio signal processing.

Jianhua Tao (Senior Member, IEEE) received the MS degree from Nanjing University, Nanjing, China, in 1996 and the PhD degree from Tsinghua University, Beijing, China, in 2001. He is currently a professor with Department of Automation, Tsinghua University, Beijing, China. He has authored or coauthored more than eighty papers on major journals and proceedings. His current research interests include speech recognition, speech synthesis and coding methods, human–computer interaction, multimedia information processing, and pattern recognition. He is the chair or program committee member for several major conferences, including ICPR, ACII, ICMI, ISCSLP, etc. He is also the steering committee member for the *IEEE Transactions on Affective Computing*, an associate editor for *Journal on Multimodal User Interface* and *International Journal on Synthetic Emotions*, and the deputy editor-in-chief for *Chinese Journal of Phonetics*.