UniMF: A Unified Multimodal Framework for Multimodal Sentiment Analysis in Missing Modalities and Unaligned Multimodal Sequences

Ruohong Huan, Guowei Zhong, Peng Chen, *Member, IEEE*, and Ronghua Liang, *Senior Member, IEEE*

***Abstract*—In current multimodal sentiment analysis, aligned and complete multimodal sequences are often crucial. Obtaining complete multimodal data in the real world presents various challenges, and aligning multimodal sequences often requires a significant amount of effort. Unfortunately, most multimodal sentiment analysis methods fail when dealing with missing modalities or unaligned multimodal sequences. To tackle these two challenges simultaneously in a simple and lightweight manner, we present the Unified Multimodal Framework (UniMF). The primary components of UniMF comprise two distinct modules. The first module, Translation Module, translates missing modalities using information from existing modalities. The second module, Prediction Module, uses the attention mechanism to fuse the multimodal information and generate predictions. To enhance the translation performance of the Translation Module, we introduce the Multimodal Generation Mask (MGM) and utilize it to construct the Multimodal Generation Transformer (MGT). The MGT can generate the missing modality while focusing on information from existing modalities. Furthermore, we introduce the Multimodal Understanding Transformer (MUT) in the Prediction Module, which includes the Multimodal Understanding Mask (MUM) and a unique sequence, *MultiModalSequence* (*MMSeq*), representing a unified multimodality. To assess the performance of UniMF, we perform experiments on four multimodal sentiment datasets, and UniMF attains competitive or state-of-the-art outcomes with fewer learnable parameters. Furthermore, the experimental outcomes signify that UniMF, supported by MGT and MUT - two transformers utilizing special attention mechanisms, can efficiently manage both generating task of missing modalities and understanding task of unaligned multimodal sequences.**

***Index Terms*—Attention mechanism, missing modalities, multimodal sentiment analysis, transformer, unaligned multimodal sequences.**

Manuscript received 28 November 2022; revised 2 June 2023 and 19 August 2023; accepted 29 November 2023. Date of publication 4 December 2023; date of current version 4 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62276237, 62036009, and U1909203, in part by the Basic Public Welfare Research Program of Zhejiang Province under Grant LTGY23F020006, in part by the Zhejiang Provincial Natural Science Foundation of China under Grants LDT23F0202 and LDT23F02021F02. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Ichiro Ide. *(Corresponding author: Ronghua Liang.)*

The authors are with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: huanrh@zjut.edu.cn; guoweizhong@zjut.edu.cn; chenpeng@zjut.edu.cn; rhliang@zjut.edu.cn).

The implementation of this work is publicly accessible at https://github.com/gw-zhong/UniMF.

Digital Object Identifier 10.1109/TMM.2023.3338769

I. Introduction

IN REAL-WORLD scenarios, individuals often express information using multiple modalities. For instance, during communication, we can perceive not only the spoken words but also body movements and tone levels. This multimodal information enhances our ability to interpret the intentions behind the message. Combining information from multiple modalities to judge a person's sentiments is known as multimodal sentiment analysis (MSA), which remains a critical challenge in human-computer interaction.

MSA has attracted a lot of attention, with most works focusing on multimodal fusion methods [1], [2], [3], [4], [5], [6], [7]. For instance, Bi-directional Contextual LSTM (BC-LSTM) [1] leverages multiple LSTMs [8] in a hierarchical architecture to fuse modalities. Tensor Fusion Network (TFN) [3] uses Cartesian products to model inter-modality and intra-modality dynamics. Recurrent Attended Variation Embedding Network (RAVEN) [4] calculates the verbal and non-verbal modalities' offset and adds the values to fuse the modalities. More recently, several studies based on BERT [9] or XLNet [10] models have appeared, which explore different approaches to multimodal fusion based on new dynamic word embeddings [11], [12], [13], [14], [15], [16], [17], [18]. Though they exhibit impressive performance, most multimodal fusion approaches share two significant limitations. Firstly, the input modalities must be complete, implying that there must be no absence of one or several modalities. Secondly, the input multimodal sequences must be aligned.

Real-world multimodal sequences are typically incomplete, and addressing the problem of missing modalities has become a topic of interest in recent years [19], [20], [21], [22], [23]. Multimodal Cyclic Translation Network (MCTN) [19] proposes the use of a sequence-to-sequence (seq2seq) architecture [24], [25] to achieve mutual translation between different modalities. Coupled-Translation Fusion Network (CTFN) [22] is inspired by CycleGAN [26] and performs couple learning with a cyclic consistency constraint, using CNN to achieve multimodal fusion. Transformer-based Feature Reconstruction Network (TFR-Net) [23] focuses on randomly missing tokens within modalities. These methods ensure recognition performance in the absence of one or some modalities. However, most

methods cannot handle unaligned input multimodal sequences, imposing certain restrictions.

The heterogeneity among modalities in unaligned multimodal sequences can pose significant challenges to sentiment analysis. Previous approaches align all modalities except *Language* modality at the word level, which is not only costly but also results in a certain loss of information on non-verbal modalities. Recent works have been dedicated specifically to unaligned multimodal sequences [27], [28], [29], [30], [31], [32], [33], [34]. For instance, Multimodal Transformer (MulT) [27] achieves alignment of multimodal sequences through the attention mechanism in the transformer decoder. Sparse Phased Transformer (SPT) [30] reduces input sequence length using a sparse attention mechanism and reuses cross-modal attention matrices to achieve a much lighter mode. Cross Hyper-modality Fusion Network (CHFN) [33] uses unaligned non-linguistic behavioral information to adjust the human language's representation dynamically across the entire range of discourse scales. These approaches enable end-to-end sentiment analysis in multimodal sequences. However, they still suffer from a significant limitation: they only work with complete multimodal sequences. The absence of a modality or some modalities renders most of these methods ineffective.

Recently, a few methods have emerged for modality missing processing in unaligned multimodal sequences [23], [35], [36], [37], [38], [39]. However, some of these methods require multi-stage training [35], additional data [39], or complex extraction and reconstruction frameworks [23]. Additionally, almost all of these methods require a paired encoder-decoder framework for missing modality reconstruction [36], [37], [38], leading to further network complexity.

To address both modality missing and unaligned multimodal sequences in a simpler and more lightweight manner, the primary contributions of this paper can be summarized as follows:

- We propose a Unified Multimodal Framework (UniMF) designed to address modality missing and unaligned multimodal sequences in a straightforward and efficient manner. Two main modules, the Translation Module and the Prediction Module, are included in this framework.
- We propose the Translation Module, which has the Multimodal Generation Transformer (MGT) at its core. This module integrates existing modalities' information encoding and decodes the missing modality using the Multimodal Generation Mask (MGM) attention mechanism along with the `[multi]` token.
- We introduce the Prediction Module, which consists of the Multimodal Understanding Transformer (MUT) at its core. The MUT uses the Multimodal Understanding Mask (MUM) attention mechanism to exchange information between unimodal and multimodal, which ultimately enhances the final multimodal representation. To achieve a more unified multimodal representation, we utilize a unique sequence called the *MultiModalSequence* (*MMSeq*) that is initialized randomly in the MUT, allowing it to interact with the input unimodal sequences.
- We evaluate our framework's performance by testing it on four different multimodal sentiment datasets: CMU-MOSI, CMU-MOSEI, MELD, and UR-FUNNY. Our experimental findings demonstrate that UniMF outperforms or matches the state-of-the-art (SOTA) results with less learnable parameters.

The rest of the paper is organized as follows: Section II discusses the related methods, including the works that handle missing modalities, unaligned multimodal sequences, and missing modalities under unaligned multimodal sequences. In Section III, we discuss the proposed UniMF, including the Translation Module and the Prediction Module. Section IV compares UniMF with recent SOTA approaches on four datasets, and we conduct ablation studies to confirm the effectiveness of each of our proposed modules. In Section V, we compare the effects of BERT and Glove word embeddings on model performance, visualize the MUM attention mechanisms and joint representations, and analyze modality imputation performance. Lastly, Section VI concludes the paper and presents future work prospects.

II. Related Work

Most of the current approaches for MSA are focused on multimodal fusion. BC-LSTM [1] and Gated Multimodal Embedding LSTM with Temporal Attention (GME-LSTM (A)) [40] utilize LSTM-based models to capture multimodal contextual information. Poria et al. [41] and Majumder et al. [42] adopt an early fusion technique to concatenate the features of *Audio* and *Language* modalities and use GRUs [43] to model the sentiment context. Context-aware Hierarchical Fusion (CHFusion) [44] applies an RNN-based hierarchical structure to fuse two modalities and then fuse all three modalities. Ghosal et al. [2] propose a multimodal attention block using Multi-Modal Multi-Utterance-Bi-modal Attention (MMMU-BA). Zadeh et al. [3], [45] adopt the Cartesian product and a multilevel attention mechanism called Multi-Attention Recurrent Network (MARN) to perform multimodal fusion. Memory Fusion Network (MFN) [46] is used to model each modality individually and then accomplish multimodal fusion through a gated attention mechanism. Liu et al. [47] decompose the weights into low-rank factors based on TFN and name it Low-rank Multimodal Fusion (LMF). Zadeh et al. [48] introduce the Dynamic Fusion Graph (DFG) based on MFN for multimodal information fusion in Graph Memory Fusion Network (Graph-MFN). Wang et al. [4] utilize RAVEN to compute offsets between non-verbal modalities and the *Language* modality. Tsai et al. [49] use Multimodal Factorization Model (MFM) to dissolve multimodal representations into modality-invariant and modality-specific representations for inter- and intra-modal interactions. Liang et al. [5] embed multilevel fusion processes in an LSTM called Recurrent Multistage Fusion Network (RMFN). Locally confined Modality Fusion Network (LMFN) [6] explores local and global fusions of multiple modalities to achieve a holistic understanding of information. Lastly, Mai et al. [7] propose Recurrent Decomposition Fusion Network (RDFN), which uses the gyroscope structure to learn the high-level representation of each unimodal state and individual modality-specific and cross-modality interactions using inter-modal information flow.

While these methods have shown great success in multimodal fusion, they do not adequately address missing modalities or unaligned multimodal sequences. As a result, when dealing with such situations, many methods tend to underperform.

A. Dealing With Missing Modalities

Seq2Seq2Sentiment (Seq2Seq2Sent) [20] utilizes a hierarchical seq2seq architecture to create a joint representation. MCTN [19] addresses missing modalities by inter-modal translation so that in the inference phase, only one modality is required for the final sentiment analysis. TransModality [21] introduces the Modality Fusion Cell using two transformer encoder-decoder pairs for forward and backward modality translations. Tang et al. [22] leverage the transformer encoder to enhance the network's robustness for handling missing modalities. Additionally, Adaptive Modality Distillation (AMD) [50] presents a multimodal tensor-based approach using Tucker decomposition to improve computational efficiency. Yuan et al. [51] propose the Noise Intimating-based Adversarial Training (NIAT) framework using temporal feature erasing to augment noisy instances, employing the self-attention mechanism to understand modality interactions and learn multimodal representation for both original and noisy instance pairs.

Despite the effectiveness of the aforementioned techniques to handle missing modalities or incomplete data, most of them do not account for situations where input multimodal sequences are unaligned.

B. Dealing With Unaligned Multimodal Sequences

MulT [27] is the first to introduce the transformer-based MSA framework for processing unaligned sequences using cross-modal attention. Progressive Modality Reinforcement (PMR) [28] then proposes a message hub to store multimodal information separately and uses a progressive strategy to constantly exchange information between unimodal and multimodal in the forwarding process, resulting in mutual enhancement. Additionally, a dynamic filter is applied to weigh self-attention and cross-modal attention adaptively. Modality-Invariant Cross-modal Attention (MICA) [29] introduces a novel attention mechanism that learns cross-modal interactions in a modality-invariant space to compensate for the distribution mismatch between different modalities. SPT [30] improves MulT by introducing a Sparse Phased Block (SP-Block) to sample long sequences and factorized co-attention, which reduces the number of transformer blocks. Self-supervised Multi-task Multimodal sentiment analysis network (Self-MM) [31] introduces self-supervised learning to MSA by learning a unimodal label for each modality and using it as an auxiliary task during training. CHFN [33] focuses on the impact of non-verbal behavioral information across the discourse context and proposes using unaligned multimodal sequences to adjust word representations dynamically in different non-verbal contexts. Finally, Multimodal Fusion approach for learning modality-Specific and modality-Agnostic (MFSA) [34] introduces a fusion approach to improve multimodal representations and exploit the complementarity between different modalities.

However, while the methods mentioned above have shown efficacy in processing unaligned multimodal sequences, they fail to address the issue when a particular modality is missing or unavailable.

C. Dealing With Missing Modalities Under Unaligned Multimodal Sequences

Recently, some researches have emerged that address the problem of missing modalities in unaligned multimodal sequences.

TFR-Net [23] uses a transformer-based feature reconstruction network to increase the robustness of models to random missing modalities in unaligned multimodal sequences. Missing Modality Imagination Network (MMIN) [35] trains joint multimodal representations that can predict the missing modality representation based on the available modalities in various missing modality scenarios. Tag-Assisted Transformer Encoder (TATE) [36], [37] addresses the issue of missing uncertain modalities. Ensemble-based Missing Modality Reconstruction (EMMR) [38] uses a backbone encoder-decoder network to learn joint representations with available modalities and checks semantic consistency to determine whether the missing modality is crucial to the overall sentiment polarity. Graph Complete Network (GCNet) [52] consists of two graph neural network-based modules, Speaker GNN and Temporal GNN, designed to capture speaker and temporal dependencies, respectively. The network jointly optimizes classification and reconstruction tasks to make the most extensive use of both complete and incomplete data in an end-to-end manner. Efficient Multimodal Transformer with Dual-Level Feature Restoration (EMT-DLFR) [39] includes two main modules: 1) EMT utilizes global multimodal context from each modality to interact with local unimodal features and mutually promote each other at the utterance-level, and 2) DLFR performs low-level feature reconstruction to encourage the model to learn semantic information from incomplete data.

However, some of these approaches require multi-stage training [35], additional data [39], or involve using many transformer blocks to handle intra- and inter-modal interactions [23], leading to a relatively redundant overall framework.

III. Method

This section centers on the UniMF framework, depicted in Fig. 1. For instance, considering the input as the *Language* and *Audio* modalities (with missing *Video* modality), we pass these modalities through the Translation Module to generate the *Fake Video* modality. Finally, the multimodal sequence (*Language*, *Audio*, *Fake Video*) is fed into the Prediction Module for sentiment analysis prediction.

A. Preliminaries

1) Cross-Modal Attention (CA): As suggested in the previous work [27], we can calculate cross-modal attention between

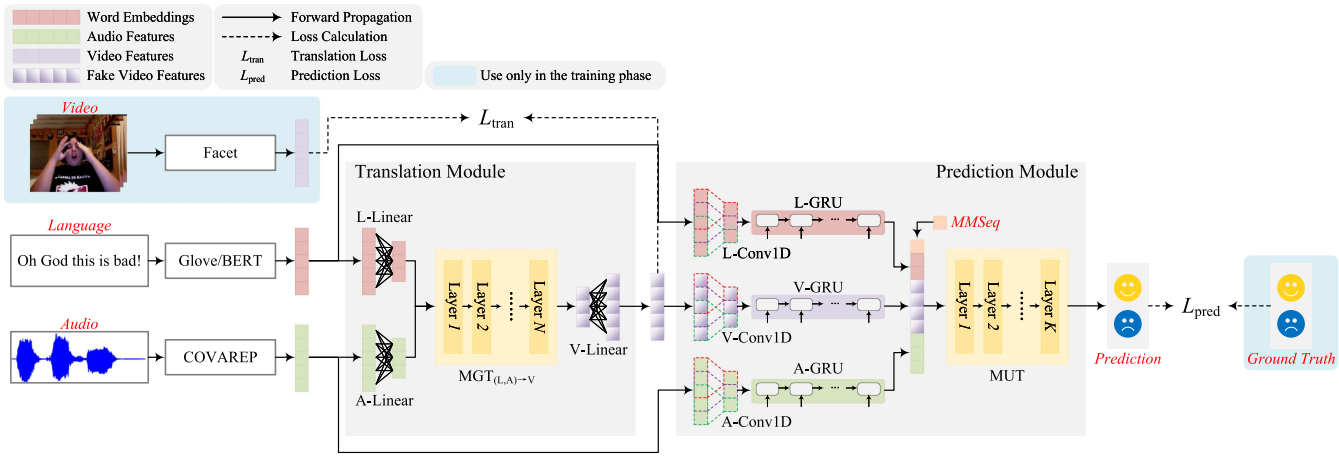


Fig. 1. UniMF workflow when the *Video* modality is missing. Specifically, UniMF mainly comprises two modules: the Translation Module and the Prediction Module. The MGT is the core of the Translation Module, and it primarily leverages the information from existing modalities to translate the missing modality (if there are no missing modalities, no translation is performed). The Prediction Module is centered around the MUT, which fuses the information from multimodal sequences whether or not they are aligned, and produces the sentiment analysis result.

modality α ($X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$) and modality β ($X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$):

$$\begin{aligned} Y_\alpha &= \mathrm{CA}_{\beta\to\alpha}(X_\alpha, X_\beta) \\ &= \mathrm{softmax}\left(\frac{Q_\alpha K_\beta^T}{\sqrt{d_k}}\right) V_\beta \\ &= \mathrm{softmax}\left(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^T X_\beta^T}{\sqrt{d_k}}\right) X_\beta W_{V_\beta} \end{aligned} \tag{1}$$

where $Y_\alpha \in \mathbb{R}^{T_\alpha \times d_v}$, $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}$, $W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$, and $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$. The cross-modal attention mechanism enables Y_α to preserve the length of modality α while being mapped into the feature space of modality β. Therefore, we can consider that cross-modal attention facilitates the flow of information from modality β to modality α.

2) Self-Attention (SA): Based on the input $X \in \mathbb{R}^{T \times d}$, the self-attention [53] can be calculated as follows:

$$\begin{aligned} Y &= \mathrm{SA}(X) \\ &= \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \\ &= \mathrm{softmax}\left(\frac{XW_Q W_K^T X^T}{\sqrt{d_k}}\right) XW_V \end{aligned} \tag{2}$$

where $Y \in \mathbb{R}^{T \times d_v}$, $W_Q \in \mathbb{R}^{d \times d_k}$, $W_K \in \mathbb{R}^{d \times d_k}$, and $W_V \in \mathbb{R}^{d \times d_v}$.

B. Translation Module

1) Multimodal Generation Mask (MGM): We propose the Multimodal Generation Mask (MGM) attention mechanism to efficiently unify the encoder-decoder pair. This mechanism introduces a special token, [multi], which acts as the starting token for generating missing modality during the inference phase. Unlike the [SOS] token in seq2seq architecture, the [multi] token includes contextual information of the existing modalities through the MGM attention mechanism. Fig. 2 illustrates the MGM attention mechanism for the case of inputting *Language* and *Audio* modalities with the *Video* modality missing, i.e., $\mathrm{MGM}_{(\mathrm{L,\,A})\to\mathrm{V}}$.

As illustrated in Fig. 2, the MGM is produced through the addition of a MASK^G specific to the modality translation direction into the self-attention matrix. In the encoder component, which contains the existing modalities and the [multi] token, the current modality can perform self-attention and transfer its information to the [multi] token. On the other hand, in the decoder component, which contains the [multi] token and the missing modality, the token of the missing modality concentrates on itself as well as the tokens on its left side. In this context, the [multi] token serves as a pivotal bridge connecting both the encoder and the decoder components, actively engaging with and contributing to both segments of the MGM.

In implementing MASK^Gs, we opt to assign the value $-inf$ to the portions to be masked and 0 to the portions to be unmasked. Thus, after performing the subsequent softmax operation, the attention weights of the masked area become 0, and the attention weights of the unmasked portions remain constant.

Mathematically, the MGM attention mechanism can be calculated as follows:

$$\begin{aligned} Y &= \mathrm{MGM}(X) \\ &= \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \mathrm{MASK}^G\right) V \\ &= \mathrm{softmax}\left(\frac{XW_Q W_K^T X^T}{\sqrt{d_k}} + \mathrm{MASK}^G\right) XW_V \end{aligned}$$

$$\mathrm{MASK}_{ij}^G = \begin{cases} 0, & \text{allow to attend} \\ -\infty, & \text{prevent from attending} \end{cases} \tag{3}$$

where $Y \in \mathbb{R}^{T \times d_v}$, $W_Q \in \mathbb{R}^{d \times d_k}$, $W_K \in \mathbb{R}^{d \times d_k}$, and $W_V \in \mathbb{R}^{d \times d_v}$. The attention mask matrix $\mathrm{MASK}^G \in \mathbb{R}^{T \times T}$ is used to

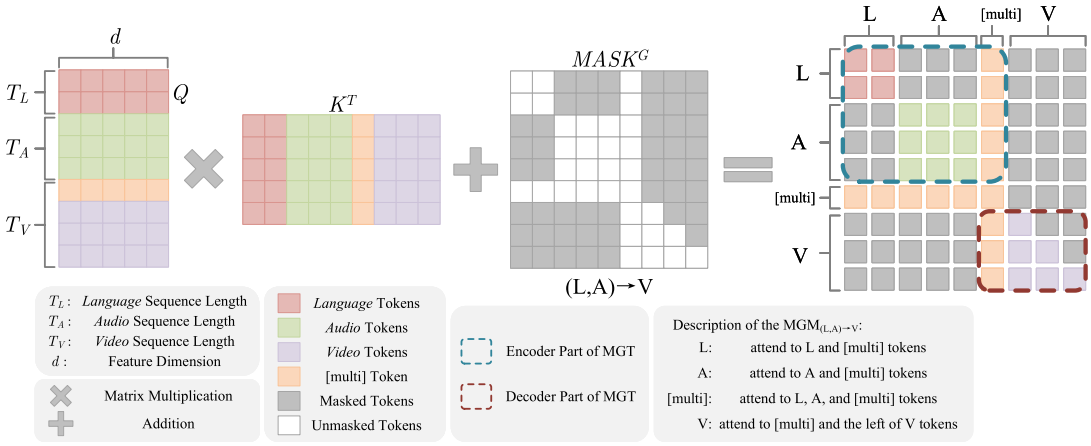


Fig. 2. Multimodal Generation Mask (MGM) for the direction (L,A)→V (MGM$_{(\mathrm{L,\,A})\rightarrow\mathrm{V}}$).

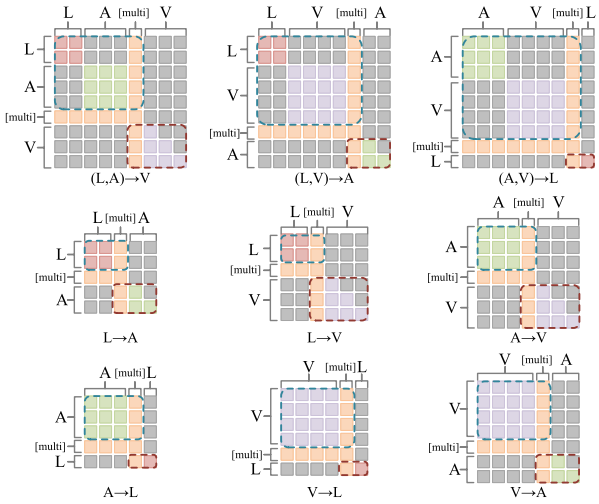


Fig. 3. MGM attention mechanism for different translation directions.

decide which pairs of tokens can attend to each other. The subscript ij of MASK^G_{ij} refers to the ith row and jth column of the matrix, while the superscript G specifies that this is the attention mask matrix employed for missing modality generation.

We create various MASK^Gs for the attention matrix, enabling us to obtain multiple MGMs across different directions of modality translation, as depicted in Fig. 3.

2) Multimodal Generation Transformer (MGT): The Multimodal Generation Transformer (MGT) is primarily composed of MGM. Specifically, for the input $X \in \mathbb{R}^{T\times d}$, the mth layer of the MGT can be calculated as follows:

$$\begin{aligned} X^{[0]} &= X \\ \hat{X}^{[m]} &= \mathrm{MGM}\left(\mathrm{LN}\left(X^{[m-1]}\right)\right) + \mathrm{LN}\left(X^{[m-1]}\right) \\ X^{[m]} &= \mathrm{FFN}\left(\mathrm{LN}\left(\hat{X}^{[m]}\right)\right) + \mathrm{LN}\left(\hat{X}^{[m]}\right) \end{aligned} \tag{4}$$

where LN denotes layer normalization [54], and FFN refers to the feed-forward network with ReLU as the activation function.

3) Forward Propagation: To simplify the explanation, we illustrate the proposed modality translation method with the input modalities *Language* (L) and *Audio* (A), and the missing modality *Video* (V) as an example. For the input sequence: $X_L = [x_{L_1}, \ldots, x_{L_{T_L}}] \in \mathbb{R}^{T_L\times d_L}$, $X_A = [x_{A_1}, \ldots, x_{A_{T_A}}] \in \mathbb{R}^{T_A\times d_A}$, $X_V = [x_{V_1}, \ldots, x_{V_{T_V}}] \in \mathbb{R}^{T_V\times d_V}$, we initially map them to the same feature dimension d so that the sequences can be concatenated in the temporal dimension:

$$\bar{X}_{\{L,A,V\}} = \mathrm{ReLU}\left(X_{\{L,A,V\}}W_{\{L,A,V\}} + b_{\{L,A,V\}}\right) \tag{5}$$

where $W_{\{L,A,V\}} \in \mathbb{R}^{d_{\{L,A,V\}}\times d}$ and $b_{\{L,A,V\}} \in \mathbb{R}^{T_{\{L,A,V\}}\times d}$. To generate the target modality *Video*, we perform a sequence modification that involves removing the last token of the input *Video* sequence and embedding the special token `[multi]` at the beginning of the sequence with a random initialization. By doing this, we complete the shifted right operation at the input as shown below:

$$\bar{X}'_V = [\texttt{multi}, \bar{x}_{V_1}, \ldots, \bar{x}_{V_{T_V-1}}] \in \mathbb{R}^{T_V\times d} \tag{6}$$

Then, position embedding (PE) and modal-type embedding (ME) are added to each modality sequence, respectively:

$$\begin{aligned} Z_L &= \bar{X}_L + \mathrm{PE} + \mathrm{ME} = [z_{L_1}, \ldots, z_{L_{T_L}}] \\ Z_A &= \bar{X}_A + \mathrm{PE} + \mathrm{ME} = [z_{A_1}, \ldots, z_{A_{T_A}}] \\ Z_V &= \bar{X}'_V + \mathrm{PE} + \mathrm{ME} = [z_M, z_{V_1}, \ldots, z_{V_{T_V-1}}] \end{aligned} \tag{7}$$

We set the ME of *Language* to 0, *Audio* to 1, and *Video* to 2 in this paper. Furthermore, we set both the PE and the ME as learnable parameters in our experiments. Following the addition of the PE and ME, we concatenate the sequences of each modality along the temporal dimension and pass them to the MGT composed of N layers:

$$\begin{aligned} Z^{[0]}_{\{L,A,V\}} &= Z_{\{L,A,V\}} \\ Z^{[0]} &= \mathrm{concat}\left(Z^{[0]}_L, Z^{[0]}_A, Z^{[0]}_V\right) \\ Z^{[n]} &= \mathrm{MGT}_{(L,A)\rightarrow V}\left(Z^{[n-1]}\right) \end{aligned} \tag{8}$$

where $n = 1, \ldots, N$. For each token of the *Language* modality, the following attention computation is performed in each layer of the MGT. We omit the coefficients in front of CA or SA in the following equation for simplicity, as they sum to 1 due to the softmax operations present in the attention mechanism:

$$Y_L = \mathrm{CA}_{[\texttt{multi}]\to L}\left(Z_L, [\texttt{multi}]\right) + \mathrm{SA}\left(Z_L\right) \tag{9}$$

This setup enables the *Language* sequence to receive information from the multimodal token [multi] while focusing on itself. Similarly, there is an information interaction for each token of the *Audio* modality:

$$Y_A = \mathrm{CA}_{[\texttt{multi}]\to A}\left(Z_A, [\texttt{multi}]\right) + \mathrm{SA}\left(Z_A\right) \tag{10}$$

For the multimodal token [multi], its main function is to receive messages from other modalities:

$$\begin{aligned} Y_{[\texttt{multi}]} = &\ \mathrm{CA}_{L\to[\texttt{multi}]}\left([\texttt{multi}], Z_L\right) \\ &+ \mathrm{CA}_{A\to[\texttt{multi}]}\left([\texttt{multi}],\ Z_A\right) \\ &+ \mathrm{SA}\left([\texttt{multi}]\right) \end{aligned} \tag{11}$$

We implement a standard autoregressive masked self-attention mechanism for the missing *Video* modality. This means that each token of the *Video* sequence can only attend to itself and the preceding tokens on the left side. By exchanging information across modalities, the [multi] token effectively captures all necessary information and serves as a better contextual representation for generating the first token of the *Video* modality during inference.

Finally, we extract the $Z^{[N]}$ output corresponding to the *Video* modality and denote it as $Z_V^{[N]} \in \mathbb{R}^{T_V \times d}$. After resizing $Z_V^{[N]}$ to its original dimensions, we align it with the original target sequence X_V using mean square error (MSE) loss in the form of a teacher-forcing mechanism:

$$\begin{aligned} \hat{X}_V &= Z_V^{[N]} W^O + b^O \\ \mathcal{L}_{(L,A)\to V} &= \left(\hat{X}_V - X_V\right)^2 \end{aligned} \tag{12}$$

where $W^O \in \mathbb{R}^{d\times d_V}$ and $b^O \in \mathbb{R}^{T_V \times d_V}$. During the inference phase, we begin with the [multi] token and generate the complete *Fake Video* sequence $\hat{X}_V$ token by token:

$$\begin{aligned} p\left(X_V\right) &= \prod_{l=1}^{T_V} p\left(x_{V_l} | [\texttt{multi}], x_{V_1}, \ldots, x_{V_{l-1}}\right) \\ \hat{X}_V &= \arg\max_{X_V} p\left(X_V | X_L, X_A\right) \end{aligned} \tag{13}$$

After obtaining $\hat{X}_V$, it can be fed into the next Prediction Module for further sentiment prediction.

It is worth noting that when two modalities are missing, we continue to apply the one-to-one translation approach and use two separate MGTs to translate the different modalities. This decision is based on the assumption that it is challenging to translate two modalities in one transformer with a unified coding architecture. As an example, when the input modality is *Language* and the *Audio* and *Video* modalities are missing, we optimize the $\mathcal{L}_{L\to(A,V)}$:

$$\begin{aligned} \mathcal{L}_{L\to A} &= \left(\hat{X}_A - X_A\right)^2 \\ \mathcal{L}_{L\to V} &= \left(\hat{X}_V - X_V\right)^2 \\ \mathcal{L}_{L\to(A,V)} &= \mathcal{L}_{L\to A} + \mathcal{L}_{L\to V} \end{aligned} \tag{14}$$

C. Prediction Module

1) Multimodal Understanding Mask (MUM): We introduce a novel attention mechanism called the Multimodal Understanding Mask (MUM) for facilitating intra-modal and inter-modal interactions, as depicted in Fig. 4. In addition, we propose a special sequence named *MultiModalSequence* (*MMSeq* (M)), which is initialized randomly with $X_M = [x_{M_1}, \ldots, x_{M_{T_M}}] \in \mathbb{R}^{T_M \times d}$, to obtain a unified multimodal representation.

The MASK^U matrix is specifically designed to enable effective information exchange between the unimodal and multimodal modalities. During each attention interaction, *MMSeq* receives information from all modalities including *Language*, *Audio*, and *Video*. At the same time, *Language*, *Audio*, and *Video* receive information from *MMSeq* based on the completion of self-attention. This mutual information exchange between the modalities is facilitated by the MASK^U matrix and is a critical aspect of the MUM attention mechanism.

Similarly to the MASK ^{G} s, we set the parts of the MASK^U matrix that require a mask to $-inf$, and the parts that do not require a mask to 0. This way, we ensure that only the relevant information is exchanged between the modalities and prevent redundant or irrelevant information from being incorporated during the attention mechanism.

Mathematically, the MUM attention mechanism can be calculated as follows:

$$\begin{aligned} Y &= \mathrm{MUM}(X) \\ &= \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \mathrm{MASK}^U\right) V \\ &= \mathrm{softmax}\left(\frac{XW_Q W_K^T X^T}{\sqrt{d_k}} + \mathrm{MASK}^U\right) XW_V \\ \mathrm{MASK}_{ij}^U &= \begin{cases} 0, & \text{allow to attend} \\ -\infty, & \text{prevent from attending} \end{cases} \end{aligned} \tag{15}$$

where $Y \in \mathbb{R}^{T\times d_v}$, $W_Q \in \mathbb{R}^{d\times d_k}$, $W_K \in \mathbb{R}^{d\times d_k}$, and $W_V \in \mathbb{R}^{d\times d_v}$. The attention mask matrix $\mathrm{MASK}^U \in \mathbb{R}^{T\times T}$ decides if two tokens can be attended to each other. The subscript ij of MASK_{ij}^U represents the ith row and jth column of the matrix and the superscript U signifies that it is a multimodal sentiment understanding attention mask matrix.

2) Multimodal Understanding Transformer (MUT): Similar to the MGT utilized in the Translation Module, the MUT at layer m, given an input $X \in \mathbb{R}^{T\times d}$, is defined by the following equations:

$$\begin{aligned} X^{[0]} &= X \\ \hat{X}^{[m]} &= \mathrm{MUM}\left(\mathrm{LN}\left(X^{[m-1]}\right)\right) + \mathrm{LN}\left(X^{[m-1]}\right) \end{aligned}$$

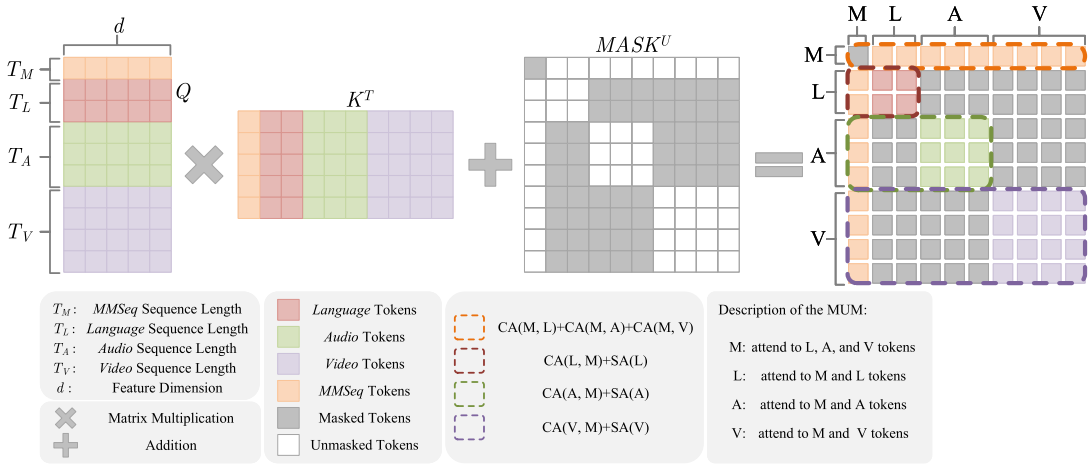


Fig. 4. Multimodal Understanding Mask (MUM).

$$X^{[m]} = \text{FFN}\left(\text{LN}\left(\hat{X}^{[m]}\right)\right) + \text{LN}\left(\hat{X}^{[m]}\right) \tag{16}$$

3) Forward Propagation: In the Prediction Module, our first step is to take in three input modalities: $X_L \in \mathbb{R}^{T_L \times d_L}$, $X_A \in \mathbb{R}^{T_A \times d_A}$, and $X_V \in \mathbb{R}^{T_V \times d_V}$. (Note: The missing modalities are generated by the Translation module, and the specially labeled fake sequences are no longer applicable in this section.) To enable attention computation, we apply 1D temporal convolution layers to map the different modalities into the same embedding space and focus on adjacent tokens:

$$\bar{X}_{\{L,A,V\}} = \text{Conv1D}\left(X_{\{L,A,V\}}, k_{\{L,A,V\}}\right) \tag{17}$$

The Transformer in isolation lacks inductive bias, which means it requires a substantial amount of data to improve its representation learning. Thus, to address these shortcomings, we propose modeling multimodal sequences using Gated Recurrent Units (GRUs) first [43]:

$$H_{\{L,A,V\}} = \text{GRU}\left(\bar{X}_{\{L,A,V\}}\right) \tag{18}$$

In a similar manner to the Translation Module, we add positional encoding (PE) and modality encoding (ME) to each of the four sequences ($H_{\{L,A,V\}}$ and X_M):

$$\begin{aligned} Z_L &= H_L + \text{PE} + \text{ME} = [z_{L_1}, \ldots, z_{L_{T_L}}] \\ Z_A &= H_A + \text{PE} + \text{ME} = [z_{A_1}, \ldots, z_{A_{T_A}}] \\ Z_V &= H_V + \text{PE} + \text{ME} = [z_{V_1}, \ldots, z_{V_{T_V}}] \\ Z_M &= X_M + \text{PE} + \text{ME} = [z_{M_1}, \ldots, z_{M_{T_M}}] \end{aligned} \tag{19}$$

This step involves assigning a unique ME value to each modality to differentiate between multimodal and unimodal sequences. Specifically, we set ME values to 0, 1, 2, and 3 for *Language, Audio, Video*, and *MMSeq*, respectively. The ME values for *Language, Audio,* and *Video* are the same as in the Translation Module - 0, 1, and 2, respectively. To realize the information exchange between unimodal and multimodal, we concatenate the four sequences and then feed them into the MUT composed of K layers:

$$\begin{aligned} Z^{[0]}_{\{M,L,A,V\}} &= Z_{\{M,LA,V\}} \\ Z^{[0]} &= \text{concat}\left(Z^{[0]}_M, Z^{[0]}_L, Z^{[0]}_A, Z^{[0]}_V\right) \\ Z^{[k]} &= \text{MUT}\left(Z^{[k-1]}\right) \end{aligned} \tag{20}$$

where $k = 1, \ldots, K$. The MUM attention mechanism enables the exchange of messages between different modalities for each layer of the MUT, similar to the Translation Module. For the sake of simplicity, the coefficients before CA or SA are omitted. For each $Y_*(* \in \{M, L, A, V\})$, the sum of coefficients is always 1. The flows of messages can be represented as follows:

$$\begin{aligned} Y_M &= \text{CA}_{L\to M}\left(Z_M, Z_L\right) \\ &\quad + \text{CA}_{A\to M}\left(Z_M, Z_A\right) \\ &\quad + \text{CA}_{V\to M}\left(Z_M, Z_V\right) \\ Y_L &= \text{CA}_{M\to L}\left(Z_L, Z_M\right) + \text{SA}\left(Z_L\right) \\ Y_A &= \text{CA}_{M\to A}\left(Z_A, Z_M\right) + \text{SA}\left(Z_A\right) \\ Y_V &= \text{CA}_{M\to V}\left(Z_V, Z_M\right) + \text{SA}\left(Z_V\right) \end{aligned} \tag{21}$$

In addition to receiving information from the *Language*, *Audio*, and *Video* modalities, the special sequence *MMSeq* concurrently fuses them. For each individual modality, messages are collected from the *MMSeq* sequence, and each modality's base representation is enriched through self-attention. During each layer of the MUT, the information cycles between the multimodal sequence M and the unimodal sequences L, A, and V, continuously completing the interactions both within and across modalities, resulting in a more effective multimodal fusion and a robust multimodal representation. We denote the multimodal representation in $Z^{[K]}$ as $Z^{[K]}_M \in \mathbb{R}^{T_M \times d}$, which is passed through a fully connected layer to obtain the final prediction.

TABLE I
DIVISION OF DATASETS USED IN THE EXPERIMENTS AND THE SEQUENCE LENGTH OF MULTIMODAL SEQUENCES FOR EACH DATASET

Datasets	Training	Validation	Test	Sequence Length		
				L	A	V
CMU-MOSI [55]	1284 (1284)	229 (229)	686 (686)	50 (50)	50 (375)	50 (500)
CMU-MOSEI [48]	16265 (16326)	1869 (1871)	4643 (4659)	50 (50)	50 (500)	50 (500)
MELD [41]	1038 (-)	114 (-)	280 (-)	33 (-)	33 (-)	- (-)
UR-FUNNY [56]	10598 (-)	2626 (-)	3290 (-)	360 (-)	360 (-)	360 (-)

The unaligned setting is indicated inside the brackets, and datasets without this setting are denoted with A '-'

In addition, it is important to note that L1 loss is used for training in the prediction module, as expressed below:

$$\mathcal{L}_{\text{pred}} = |\hat{y} - y| \tag{22}$$

D. Loss Functions

Finally, we combine the loss functions of the translation module and the prediction module to achieve the ultimate optimization objective of UniMF, as expressed below:

$$\mathcal{L} = \mathcal{L}_{\text{tran}} + \mathcal{L}_{\text{pred}} \tag{23}$$

where $\mathcal{L}_{\text{tran}} \in \{\mathcal{L}_{(L,A)\to V}, \mathcal{L}_{(L,V)\to A}, \mathcal{L}_{(A,V)\to L}, \mathcal{L}_{L\to(A,V)}, \mathcal{L}_{A\to(L,V)}, \mathcal{L}_{V\to(L,A)}\}$.

IV. EXPERIMENTS

A. Datasets

1) Statistics of Datasets: The **CMU-MOSI** [55] dataset is derived from video blogs on YouTube and consists of 93 videos from 89 individual speakers, including 41 female speakers and 48 male speakers. The dataset is selected from 2199 video clips, and the sentiment of each video is labeled as a real value in the range of [−3, 3], which represents extremely negative to extremely positive sentiments.

CMU-MOSEI [48] is a larger dataset than CMU-MOSI and contains 22,586 video clips from 1,000 different speakers, covering 250 topic types. Each video clip is labeled with a sentiment value in the range of [−3, 3].

MELD [41] is a collection of clips taken from the television series 'Friends', comprising over 1400 dialogue pairs and 13,000 phrases. The dataset can be divided into different subsets based on the tags used, such as the Sentiment and Emotion datasets. In the Sentiment dataset, each segment is labeled with one of three emotions: positive, neutral, or negative. Meanwhile, in the Emotion dataset, each segment is labeled with one of seven emotions: neutral, joy, sadness, anger, surprise, fear, or disgust. Unlike other multimodal datasets, MELD includes only two modalities, namely *Language* and *Audio*.

The **UR-FUNNY** [56] dataset, which collects 1,866 videos of TED Talks from 1,741 speakers covering 417 topics, consists of 16,514 video clips. The team indicated humorous sentences using 'laughter markers' to mark the sentences that triggered laughter from the audience for indicating humor. The dataset is evenly split, with half being humorous, and half being non-humorous.

For each dataset, the statistics are shown in Table I.

2) Evaluation Metrics: For the CMU-MOSI and CMU-MOSEI datasets, the evaluation metrics employed are binary classification accuracy for positive and negative sentiment, denoted as Acc, and F1 score, denoted as F1. On the other hand, for the MELD and the UR-FUNNY dataset, the evaluation metrics are binary classification accuracy for each emotion.

TABLE II
HYPERPARAMETER OPTIMIZATION SETTINGS

	Range	Step Size	Distribution
MGT Layers	[1, 8]	1	-
ReLU Dropout	[0.0, 0.5]	-	Uniform
Attention Dropout	[0.0, 0.5]	-	Uniform
Language Embedding Dropout	[0.0, 0.5]	-	Uniform
Output Dropout	[0.0, 0.5]	-	Uniform
Kernel Size	[1, 5]	2	-
MUT Layers	[1, 8]	1	-

B. Experimental Details

1) Features. Language: For each of the four datasets (CMU-MOSI, CMU-MOSEI, UR-FUNNY, and MELD), we utilized the pre-trained Glove model [58] to extract word embeddings with a dimensionality of 300. Additionally, we followed the approach of previous research [22] for MELD, feeding the Glove word embeddings into a 1D-CNN [59] to obtain 600-dimensional features. To gain insights from the BERT-based comparison, we used the pre-trained $\text{BERT}_{\text{BASE}}$ model with 768 dimensions to extract word embeddings.

Audio: To extract audio features with 5 dimensions in the CMU-MOSI dataset, 74 dimensions in the CMU-MOSEI dataset, and 81 dimensions in the UR-FUNNY dataset we utilized COVAREP [60]. Whereas, for the MELD dataset, we use openSMILE [61] to extract audio features with 600 dimensions in the Sentiment sub-dataset and 300 dimensions in the Emotion sub-dataset.

Video: To extract video features from the CMU-MOSI and CMU-MOSEI datasets, we leverage the Facet tool [1]. The dimension of the video features extracted is 20 for CMU-MOSI and 35 for CMU-MOSEI. For UR-FUNNY, we use OpenFace [62] to extract video features, resulting in 75 dimensions.

2) Hyperparameters: To select optimal parameters, we utilize Optuna's [63] Tree-structured Parzen Estimator (TPE) sampler in our experiments. The specific parameter optimization settings are shown in Table II. Certain parameters are fixed during the training phase, as shown in Table III. The plateau learning rate scheduler is applied to optimize the learning rate in our experiments. This scheduler reduces the learning rate by a factor of 0.1 whenever the validation loss does not decrease for 20 consecutive iterations.

[1] iMotions 2017. https://imotions.com/

TABLE III
FIXED HYPERPARAMETERS IN EXPERIMENTS

Settings	CMU-MOSI	CMU-MOSEI	MELD	UR-FUNNY
Batch Size	128 (64)	16 (16)	128 (-)	16 (-)
Epochs	100 (100)	20 (20)	100 (-)	20 (-)
Gradient Clip	0.8 (0.8)	1.0 (1.0)	0.8 (-)	1.0 (-)
Initial Learning Rate	1e-3 (1e-3)			
Hidden Unit Size d	32 (32)			
Attention Heads	8 (8)			
Optimizer	Adam [57]			

The unaligned setting is indicated inside the brackets, and datasets without this setting are denoted with A '-'

For our BERT experiments, we use the settings of Self-MM [31] for consistency and fine-tuned the learning rate of BERT to 5e-5. The weight decay of BERT is set to 1e-3, and the layer normalization's bias and weights are not decayed. For CMU-MOSEI, we set the number of epochs to 10 and the batch size to 128 to expedite training (the rest of the dataset's settings are the same as the Glove version of the experiments). Our experiments are conducted using the PyTorch deep learning framework on NVIDIA GeForce RTX 2080Ti×4. Additionally, we set the random seed to 1111 for all experiments.

C. Experiments on Missing Modalities

UniMF demonstrates competitive or SOTA performance in addressing various modality deficiencies across four word-aligned datasets: CMU-MOSI, MELD, CMU-MOSEI, and UR-FUNNY, as presented in Tables IV–VII. In accordance with Tang et al. [22], we evaluate our UniMF on the CMU-MOSI and MELD datasets. For our benchmark comparisons, we consider several models, namely, **GME-LSTM (A)** [40], **CHFusion** [44], **MMMU-BA** [2], **Seq2Seq2Sent** [20], **MCTN** [19], **TransModality** [21], and **CTFN** [22]. Additionally, we reproduce **BC-LSTM** [1], **MELD-based** [41], **MMIN** [35], **TFR-Net** [23], and **EMT-DLFR** [39] using the same experimental setup. On the other hand, for the UR-FUNNY and CMU-MOSEI datasets, we compare our UniMF against five reproduced models, namely, **BC-LSTM** [1], **MELD-based** [41], **MMIN** [35], **TFR-Net** [23], and **EMT-DLFR** [39].

In particular, Table IV shows that UniMF significantly outperforms CTFN [22], the current SOTA method, in the majority of missing modality scenarios on the CMU-MOSI dataset, except when the input patterns are solely *Audio* or both *Audio* and *Video*; in these cases, it is lower than CTFN by 1.83% and 2.28%, respectively. Additionally, UniMF has a significantly lower average number of parameters of only 148K.

UniMF is also capable of achieving comparable or even superior performance with only 187K parameters on the MELD dataset, as shown in Table V. Specifically, UniMF achieves a 0.77% higher accuracy than the MELD-based method [41] on the MELD (Sentiment) dataset when using only the *Language* modality as input, and a 0.35% higher accuracy than GME-LSTM (A) [40] when using only the *Audio* modality as input. When using the full modality as input, UniMF performs as equally well as the SOTA method CTFN. On the MELD (Emotion) dataset, UniMF achieves a 0.2% higher accuracy than GME-LSTM (A) when using only the *Language* modality as input, but is 0.04% and 1.41% lower in accuracy than MMIN [35] and TransModality [21], respectively, when using only the *Audio* modality as input or full modality as input.

We believe that CTFN achieves more accurate non-verbal to verbal modality translation by employing bidirectional translation and cyclic consistency constraints during training. However, this approach has its drawbacks. Firstly, for every two modalities, CTFN requires the addition of a translation direction. Secondly, despite the transformer encoder being a key contribution in making the model lighter, CTFN cannot handle unaligned multimodal sequences. UniMF addresses the challenge of modality missing under unaligned multimodal sequences while maintaining a lightweight design. To achieve this, UniMF discards the bidirectional translation and cyclic consistency constraints and integrates encoder-decoder through the attention mechanism of MGM. Despite UniMF losing approximately 2% performance compared to CTFN when *Language* modality is missing on the CMU-MOSI dataset, it outperforms CTFN in all other modality missing cases. We consider this trade-off acceptable, as UniMF is still able to cope with unaligned multimodal sequences, and has approximately 42.3% less parameters than CTFN (as shown in Table V).

Table VI shows that on the CMU-MOSEI dataset, UniMF achieves SOTA performance in all input patterns except for (*Language*, *Audio*), where it falls 0.4% short of MMIN. However, UniMF has approximately 91.6% less parameters on average than MMIN, making it a more lightweight and efficient option.

UniMF achieves SOTA results on all input patterns except for (*Audio*, *Video*) on the UR-FUNNY dataset, where it falls 0.98% short of MMIN, as shown in Table VII. Nonetheless, UniMF still has approximately 90.3% fewer parameters than MMIN on average, which makes it a more efficient alternative.

In general, a multitude of experiments have demonstrated that UniMF is able to perform the modality missing task effectively while retaining a low number of parameters, courtesy of the encoder-decoder integrated MGT design.

D. Experiments on Unaligned Multimodal Sequences

To further evaluate the performance of the UniMF on unaligned multimodal sequences, experiments are conducted on the CMU-MOSI (Unaligned) and CMU-MOSEI (Unaligned) datasets. The present study follows the experimental setup described in Cheng et al. [30] and compares our proposed UniMF primarily with five other models in the field, namely, **LF-LSTM** [27], **RAVEN** [4], **MCTN** [19], **MulT** [27], and **SPT** [30]. Additionally, we reproduce **MISA** [12], **Self-MM** [31], **MMIM** [32], **MMIN** [35], **TFR-Net** [23], and **EMT-DLFR** [39] using the same experimental setup for comparison.

As depicted in Table VIII, using Glove-based word embeddings, UniMF yields 1.88% and 1.84% higher accuracy and F1-score on the CMU-MOSI dataset, respectively, compared to the SOTA method SPT [30]. Similarly, on the CMU-MOSEI dataset, UniMF achieves 0.1% higher accuracy than SPT. Although the F1-score of UniMF on the same dataset was 0.3%

TABLE IV
COMPARISON OF UNIMF WITH VARIOUS SOTA MODELS ON THE CMU-MOSI DATASET

Models◇	CMU-MOSI (Word Aligned)							
	Uni-modality			Bi-modality			Tri-modality	
	(L)	(A)	(V)	(A, V)	(L, V)	(L, A)	(L, A, V)	Size*
AMD [50]	71.80	56.70	55.20	55.70	69.40	70.70	74.00	-
GME-LSTM (A) [40]	71.30	55.40	52.30	52.90	74.30	73.50	76.50	-
CHFusion [44]	-	-	-	54.49	74.77	78.54	76.51	-
MMMU-BA [2]	-	-	-	57.45	80.85	79.92	81.25	-
Seq2Seq2Sent [20]	-	-	-	58.00	67.00	66.00	70.00	-
MCTN [19]	-	-	-	53.10	76.80	76.40	79.30	-
TransModality [21]	-	-	-	59.97	80.58	81.25	82.71	-
CTFN [22]	80.79	**61.43**	60.98	**64.48**	81.55	82.16	82.77	-
BC-LSTM† [1]	79.27	56.71	56.86	57.62	79.12	80.79	78.05	492K
MELD-based† [41]	78.20	59.15	60.52	61.13	77.44	78.20	78.20	1.17M
MMIN† [35]	72.97	56.76	54.95	59.46	72.97	72.07	74.70	1.56M
TFR-Net† [23]	77.13	49.24	55.03	47.71	78.81	78.96	76.52	880K
EMT-DLFR† [39]	76.22	53.05	45.12	49.85	73.48	76.83	73.17	1.34M
UniMF (Ours)	**82.77**	59.60	**61.89**	62.20	**82.62**	**82.77**	**83.08**	**148K**

The results reported in the table are all binary classification accuracy ◇: language encoder for all models is glove *: all sizes are derived by averaging the models' learnable parameters over the seven conditions †: reproduced from open-source code with glove word embeddings.
The best results are shown in bold and the second best results are underlined.

TABLE V
COMPARISON OF UNIMF WITH VARIOUS SOTA MODELS ON THE MELD (SENTIMENT/EMOTION) DATASET

Models◇	MELD (Word Aligned)						
	Sentiment			Emotion			
	Uni-modality		Bi-modality	Uni-modality		Bi-modality	
	(L)	(A)	(L, A)	(L)	(A)	(L, A)	Size*
GME-LSTM (A) [40]	65.52	52.03	66.46	59.57	49.59	60.01	-
CHFusion [44]	-	-	65.85	-	-	58.35	-
MMMU-BA [2]	-	-	65.56	-	-	60.26	-
Seq2Seq2Sent [20]	-	-	63.84	-	-	56.42	-
MCTN [19]	-	-	66.27	-	-	59.96	-
TransModality [21]	-	-	67.04	-	-	**61.95**	-
CTFN [22]	-	-	**67.82**	-	-	-	324K
BC-LSTM† [1]	65.98	50.17	66.19	55.08	44.66	55.94	824K
MELD-based† [41]	66.55	51.00	65.82	59.54	47.85	59.50	2.04M
MMIN† [35]	64.86	51.58	67.36	57.43	**50.00**	60.11	1.74M
TFR-Net† [23]	64.79	51.92	66.44	56.63	48.12	58.24	501K
EMT-DLFR† [39]	64.52	50.42	67.16	58.81	48.12	58.66	2.04M
UniMF (Ours)	**67.32**	**52.38**	**67.82**	**59.77**	49.96	60.54	**187K**

The results reported in the table are all binary classification accuracy ◇: language encoder for all models is glove *: all sizes are derived by averaging the models' learnable parameters over the six conditions †: reproduced from open-source code with glove word embeddings.
The best results are shown in bold and the second best results are underlined.

TABLE VI
COMPARISON OF UNIMF WITH VARIOUS SOTA MODELS ON THE CMU-MOSEI DATASET

Models◇	CMU-MOSEI (Word Aligned)							
	Uni-modality			Bi-modality			Tri-modality	
	(L)	(A)	(V)	(A, V)	(L, V)	(L, A)	(L, A, V)	Size*
BC-LSTM† [1]	80.66	64.06	64.41	65.00	80.55	80.80	80.77	530K
MELD-based† [41]	80.50	61.27	64.20	64.89	80.08	79.31	79.78	1.27M
MMIN† [35]	80.91	56.08	54.22	57.26	79.70	**81.59**	79.92	1.56M
TFR-Net† [23]	76.82	62.85	63.67	62.85	79.56	78.12	79.23	899K
EMT-DLFR† [39]	77.38	61.69	63.81	61.24	77.10	78.31	73.12	1.20M
UniMF (Ours)	**81.49**	**64.31**	**65.64**	**66.08**	**82.10**	81.19	**82.73**	**131K**

The results reported in the table are all binary classification accuracy ◇: language encoder for all models is glove *: all sizes are derived by averaging the models' learnable parameters over the seven conditions †: reproduced from open-source code with glove word embeddings.
The best results are shown in bold and the second best results are underlined.

lower than SPT, the parameters of UniMF are approximately 47.4% lower than SPT. This can be mainly attributed to the efficient information interaction enabled by the MUM attention mechanism in the UniMF. It allows the introduced special sequence, *MMSeq*, to focus effectively on the specific information of the diverse modalities, thereby producing a better unified multimodal representation, ultimately contributing to more accurate sentiment predictions.

TABLE VII
COMPARISON OF UNIMF WITH VARIOUS SOTA MODELS ON THE UR-FUNNY DATASET

Models◇	UR-FUNNY (Word Aligned)							
	Uni-modality			Bi-modality			Tri-modality	
	(L)	(A)	(V)	(A, V)	(L, V)	(L, A)	(L, A, V)	Size*
BC-LSTM† [1]	64.29	50.70	58.55	60.76	66.27	64.99	66.20	687K
MELD-based† [41]	64.69	60.26	59.56	60.76	66.00	65.69	66.20	1.68M
MMIN† [35]	62.05	62.05	57.23	**62.65**	63.03	63.03	65.79	1.61M
TFR-Net† [23]	61.67	49.30	57.14	53.52	61.57	61.47	63.18	863K
EMT-DLFR† [39]	63.88	54.23	59.56	59.26	65.59	64.89	65.49	1.28M
UniMF (Ours)	**65.39**	**62.27**	**60.46**	61.67	**67.00**	**66.10**	**69.70**	**156K**

The results reported in the table are all binary classification accuracy ◇: language encoder for all models is glove *: all sizes are derived by averaging the models' learnable parameters over the seven conditions †: reproduced from open-source code with glove word embeddings
The best results are shown in bold and the second best results are underlined.

TABLE VIII
COMPARISON OF UNIMF WITH VARIOUS SOTA MODELS ON CMU-MOSI (UNALIGNED) AND CMU-MOSEI (UNALIGNED) DATASETS

	CMU-MOSI (Unaligned)		CMU-MOSEI (Unaligned)		
Models◇	Acc	F1	Acc	F1	Size
LF-LSTM [27]	77.60	77.80	77.50	78.20	-
RAVEN [4]	72.70	73.10	75.40	75.70	-
MCTN [19]	75.90	76.40	79.30	79.30	-
MulT [27]	81.10	81.00	81.60	81.60	1.56M
SPT [30]	81.20	81.30	82.40	**82.70**	154K
MISA† [12]	75.46	75.40	78.67	78.24	1.23M
Self-MM† [31]	75.46	75.61	77.35	76.96	88K
MMIM† [32]	69.82	69.97	70.86	71.35	138K
MMIN† [35]	75.76	75.82	80.63	80.46	743K
TFR-Net† [23]	78.35	78.28	78.51	79.45	16.77M
EMT-DLFR† [39]	74.39	74.54	76.25	74.47	1.27M
UniMF (Ours)	**83.08**	**83.14**	**82.50**	82.40	**81K**

◇: language encoder for all models is glove *: sizes are all derived by averaging the models' learnable parameters over the two datasets †: reproduced from open-source code with glove word embeddings.
The best results are shown in bold and the second best results are underlined.

E. Experiments on Missing Modalities Under Unaligned Multimodal Sequences

To further evaluate the performance of UniMF under unaligned multimodal sequences with missing modalities, we conduct experiments on two unaligned datasets, CMU-MOSI and CMU-MOSEI, and compare it with three other models: **MMIN** [35], **TFR-Net** [23], and **EMT-DLFR** [39]. These models are also capable of handling missing modalities and unaligned sequences. Tables IX and X report the performances of these models on the two datasets.

UniMF achieves a very good performance on the CMU-MOSI dataset, notwithstanding that it is 2.86% lower than MMIN when the input pattern is limited to the (*Audio*) modality. However, the average number of parameters in UniMF is about 90.5% less than MMIN. While on the CMU-MOSEI dataset, UniMF achieves the best overall performance with an average number of parameters of 164K for all modality deficiencies.

These results demonstrate that UniMF is capable of effectively handling missing modalities in unaligned multimodal sequences through its lightweight architecture that leverages a seq2seq approach, with the encoder-decoder integrated MGT aiding in the processing of missing modalities, and information interaction being performed via MUT between multimodal and unimodal representations.

F. Ablation Studies

We conduct ablation experiments to study the influence of each module in UniMF on its performance. Table XI presents the results of these experiments. For the MGT module, which comprises MGM attention mechanisms and the [multi] token, we observe about a 2% degradation in performance on (L, A)→V and (L, V)→A, and about a 1% decrease in performance on (A, V)→L after replacement with a neural machine translation (NMT) transformer. Similarly, for the MUT module, which consists of MUM attention mechanisms and *MMSeq*, we observe about a 3% degradation in performance on (L, A)→V, about a 2% decrease on (L, V)→A, and about a 5% decrease on (A, V)→L after replacement with an ordinary full attention transformer. The results indicate that MGT effectively integrates the processes of encoding and decoding, leading to enhanced interaction between the encoder and decoder components. Moreover, MUT exhibits improved capability in facilitating the exchange of multimodal information and mitigating the interference caused by redundant information when contrasted with full attention mechanisms. This enhancement subsequently contributes to the overall performance enhancement of the model.

To further investigate the role of the [multi] token and *MMSeq*, we test performance after their removal. The fourth and fifth rows of Table XI show that the removal of [multi] token decreases the performance of three modality translation directions by about 1%, whereas the removal of *MMSeq* leads to about a 2% decrease in performance on (L, A)→V, about a 1% decrease on (L, V)→A, and about a 4% decrease on (A, V)→L. These findings highlight the critical importance of the [multi] token within the framework of the MGT. When exclusively employing the mean value of input features as the initial decoding token during the testing phase, there is an inevitable decline in the ultimate performance outcome. This observation provides further validation for the significance of conveying information from the pre-existing modality to the [multi] token through the application of attention mechanisms. Furthermore, the experimental outcomes underscore the significance of the distinct *MMSeq* modality introduced in the MUT. Functioning as a comprehensive representation of multimodal inputs, *MMSeq* effectively engages with diverse unimodal modalities present within the MUT. This interaction facilitates the achievement of a more robust multimodal fusion strategy, distinctly superior to the elementary concatenation of three modalities as a means of

TABLE IX
COMPARISON OF UNIMF WITH VARIOUS SOTA MODELS ON CMU-MOSI (UNALIGNED) DATASETS

Models◇	CMU-MOSI (Unaligned)							
	Uni-modality			Bi-modality			Tri-modality	
	(L)	(A)	(V)	(A, V)	(L, V)	(L, A)	(L, A, V)	Size*
MMIN† [35]	67.57	**62.16**	60.36	61.26	72.97	66.67	75.76	1.56M
TFR-Net† [23]	78.51	52.13	45.88	61.13	78.05	78.51	78.35	14.79M
EMT-DLFR† [39]	75.46	47.56	47.26	53.35	73.32	74.24	74.39	1.34M
UniMF (Ours)	**82.47**	59.30	**62.35**	**61.74**	**81.71**	**82.16**	**83.08**	**148K**

The results reported in the table are all binary classification accuracy ◇: language encoder for all models is glove †: reproduced from open-source code with glove word embeddings.
The best results are shown in bold and the second best results are underlined.

TABLE X
COMPARISON OF UNIMF WITH VARIOUS SOTA MODELS ON CMU-MOSEI (UNALIGNED) DATASETS

Models◇	CMU-MOSEI (Unaligned)							
	Uni-modality			Bi-modality			Tri-modality	
	(L)	(A)	(V)	(A, V)	(L, V)	(L, A)	(L, A, V)	Size*
MMIN† [35]	75.93	62.79	58.92	60.88	76.73	77.74	80.63	1.56M
TFR-Net† [23]	78.51	62.85	65.35	63.07	78.59	79.47	78.51	18.76M
EMT-DLFR† [39]	79.28	61.23	63.92	64.23	76.58	78.89	76.25	1.20M
UniMF (Ours)	**80.88**	**64.01**	**65.93**	**66.51**	**81.62**	**80.76**	**82.50**	**164K**

The results reported in the table are all binary classification accuracy ◇: language encoder for all models is glove †: reproduced from open-source code with glove word embeddings.
The best results are shown in bold and the second best results are underlined.

TABLE XI
ABLATION STUDIES ON CMU-MOSI (WORD ALIGNED) DATASET ◇: LANGUAGE ENCODER FOR ALL MODELS IS GLOVE

Model Design◇	Description	CMU-MOSI (Word Aligned)	
		Acc	F1
UniMF	(L, A)→V	**82.77**	**82.59**
	(L, V)→A	**82.62**	**82.48**
	(A, V)→L	**62.20**	**61.49**
w/o MGT (MGM + [multi])	(L, A)→V	80.79	80.64
	(L, V)→A	80.79	80.89
	(A, V)→L	60.98	60.99
w/o MUT (MUM + *MMSeq*)	(L, A)→V	79.12	79.22
	(L, V)→A	80.49	80.56
	(A, V)→L	56.55	56.62
w/o [multi]	(L, A)→V	81.86	81.86
	(L, V)→A	81.86	81.87
	(A, V)→L	61.13	61.22
w/o *MMSeq*	(L, A)→V	80.79	80.77
	(L, V)→A	81.10	81.14
	(A, V)→L	58.23	58.39
w/o PE	(L, A)→V	81.71	81.71
	(L, V)→A	81.71	81.68
	(A, V)→L	61.74	61.76
w/o ME	(L, A)→V	81.71	81.72
	(L, V)→A	81.10	81.00
	(A, V)→L	61.59	60.07

The best results are shown in bold.

multimodal fusion. Such concatenation, as demonstrated by the experiments, introduces a certain level of performance degradation to the model.

Moreover, we also examine the importance of PE and ME in MGT and MUT. The last two rows of Table XI demonstrate that deleting PE or ME results in about a 1% decrease in performance across the three directions. These findings highlight the considerable importance of encoding both positional information and modality-specific details within the input sequences of both the MGT and the MUT. The absence of positional information encoding would render the attentional mechanism incapable of discerning the contextual significance of identical words situated in distinct positions. Correspondingly, the omission of modality information encoding would impede the attentional mechanism's ability to differentiate between the demarcations of various modalities within the input sequences.

Our findings indicate that MGT and MUT, as well as the [multi] token and *MMSeq*, as well as PE and ME, are all essential elements of UniMF. Any missing part leads to performance degradation. Therefore, our results confirm the significance of MGT and MUT in UniMF and highlight the importance of including additional token and sequence, such as [multi] token and *MMSeq*, to consolidate modality fusion information.

V. DISCUSSIONS

A. BERT Vs. Glove

In recent years, the 'pre-train, fine-tune' paradigm has shown considerable progress in MSA methods through the emergence of large pre-trained language models. To further confirm the effectiveness of UniMF, we conduct experiments on three datasets - CMU-MOSI, CMU-MOSEI, and UR-FUNNY - by replacing Glove word embeddings with BERT embeddings.

Table XII shows the experimental results that demonstrate the superior performance of UniMF in comparison to Glove-based and BERT-based approaches respectively on the CMU-MOSI and CMU-MOSEI datasets. Our UniMF not only outperforms most Glove-based methods but also achieves SOTA performance against various BERT-based methods. Specifically, on the CMU-MOSI dataset with Glove embeddings, UniMF outperforms SOTA methods MulT [27] and SPT [30] by 0.08% and 0.28% on Acc and F1, respectively. On the CMU-MOSEI dataset with Glove embeddings, UniMF outperforms SOTA method SPT by 0.13% and 0.18% on Acc and F1, respectively. Moreover, UniMF has about 43.5% fewer parameters under both

TABLE XII
COMPARISON OF UNIMF WITH VARIOUS SOTA MODELS ON CMU-MOSI (WORD ALIGNED) AND CMU-MOSEI (WORD ALIGNED) DATASETS

	CMU-MOSI (Word Aligned)		CMU-MOSEI (Word Aligned)		
Models	Acc	F1	Acc	F1	Size*
Glove Embedding					
LF-LSTM [27]	76.80	76.70	80.60	80.60	-
RAVEN [4]	78.00	76.60	79.10	79.50	-
MCTN [19]	79.30	79.10	79.80	80.60	-
MulT [27]	83.00	82.80	82.50	82.30	1.56M
SPT [30]	82.80	82.90	82.60	82.80	154K
BC-LSTM† [1]	78.05	78.13	80.77	80.77	637K
MELD-based† [41]	78.20	78.12	79.78	79.59	1.55M
MISA† [12]	74.70	74.30	79.48	78.99	1.23M
Self-MM† [31]	72.10	72.15	79.39	78.94	88K
MMIM† [32]	73.48	73.60	73.70	74.14	138K
MMIN† [35]	74.70	74.64	79.92	79.44	743K
TFR-Net† [23]	76.52	76.39	79.23	79.82	889K
EMT-DLFR† [39]	73.17	73.33	73.12	73.54	1.27M
UniMF (Ours)	**83.08**	**83.18**	**82.73**	**82.98**	**87K**
BERT Embedding					
BERT [9]	85.20	85.20	83.90	83.90	-
MAG-BERT [11]	86.10	86.00	84.70	84.50	-
MISA [12]	83.40	83.60	85.50	85.30	-
Self-MM‡ [31]	85.98	85.95	85.17	85.30	-
MMIM‡ [32]	86.06	85.98	85.97	85.94	-
BC-LSTM† [1]	84.91	84.80	82.99	82.98	110.49M
MELD-based† [41]	85.21	85.18	82.44	81.80	112.01M
MulT† [27]	83.99	83.91	85.28	85.25	111.26M
SPT† [30]	84.15	84.17	80.35	79.60	109.66M
MMIN† [35]	79.73	79.70	83.93	84.02	110.44M
TFR-Net† [23]	84.76	84.80	84.23	84.54	110.43M
EMT-DLFR† [39]	84.76	84.77	81.65	81.89	111.70M
UniMF (Ours)	**86.28**	**86.14**	**86.19**	**86.13**	**109.62M**

*: sizes are all derived by averaging the models' learnable parameters over the two datasets ‡: results are derived from unaligned settings †: reproduced from open-source code.
The best results are shown in bold and the second best results are underlined.

TABLE XIII
COMPARISON OF UNIMF WITH VARIOUS SOTA MODELS ON UR-FUNNY DATASET

	UR-FUNNY (Word Aligned)	
Models	Acc	Size
Glove Embedding		
TFN [3]	64.71	-
LMF [47]	65.16	-
MFN [46]	65.23	-
MISA [12]	68.60	5.34M
SPT [30]	**70.00**	158K
BC-LSTM† [1]	66.20	945K
MELD-based† [41]	66.20	2.35M
MulT† [27]	63.58	943K
Self-MM† [31]	67.00	134K
MMIM† [32]	50.70	187K
MMIN† [35]	64.39	941K
TFR-Net† [23]	63.18	863K
EMT-DLFR† [39]	65.49	1.28M
UniMF (Ours)	69.70	**65K**
BERT Embedding		
TFN [3]	68.57	-
LMF [47]	67.53	-
MISA [12]	70.61	-
BC-LSTM† [1]	64.69	110.80M
MELD-based† [41]	65.90	112.82M
MulT† [27]	68.41	110.50M
SPT† [30]	67.30	**109.69M**
MAG-BERT† [11]	69.11	111.36M
Self-MM† [31]	69.42	**109.69M**
MMIM† [32]	65.29	109.82M
MMIN† [35]	68.01	110.66M
TFR-Net† [23]	67.30	110.40M
EMT-DLFR† [39]	67.40	111.69M
UniMF (Ours)	**70.82**	109.74M

†: Reproduced from open-source code.
The best results are shown in bold and the second best results are underlined.

CMU-MOSI and CMU-MOSEI datasets with Glove embeddings compared to the SOTA method SPT, which is only 87K. On the CMU-MOSI dataset with BERT embeddings, UniMF outperforms the SOTA method MAG-BERT [11] by 0.18% and 0.14% on Acc and F1, respectively. On the CMU-MOSEI dataset with BERT embeddings, UniMF improves by 0.22% and 0.19% over the SOTA method MMIM [32] on Acc and F1, respectively.

Table XIII presents the results obtained from the UR-FUNNY dataset using Glove embeddings and BERT embeddings, respectively. UniMF's performance is found to be 0.3% lower compared to the SOTA approach SPT [30], but with a significantly fewer number of parameters at only 65K, which is about 58.9% less than SPT. On the other hand, for the UR-FUNNY dataset using BERT word embeddings, UniMF outperforms the SOTA approach MISA [12] by 0.21%.

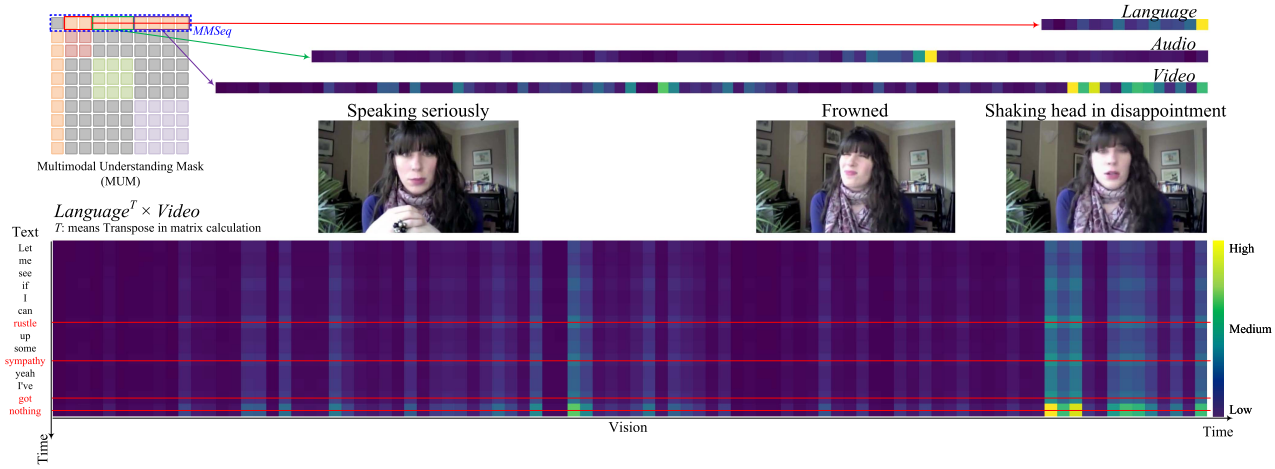
The improved performance of UniMF can be attributed to the introduction of the *MMSeq* sequence, which facilitates the exchange of information between multimodal and unimodal inputs via the MUM attention mechanism. This feature enables the creation of a better unified multimodal representation and ultimately enhances sentiment analysis performance. Prior studies [11], [15] suggest that the use of XLNet [10] as the language encoder may further enhance the model's performance. However, we believe that this performance improvement is solely due to the language encoder selection and not the framework's design. Therefore, we do not incorporate XLNet into our subsequent experiments.

B. Visualization of MUM Attention Mechanism

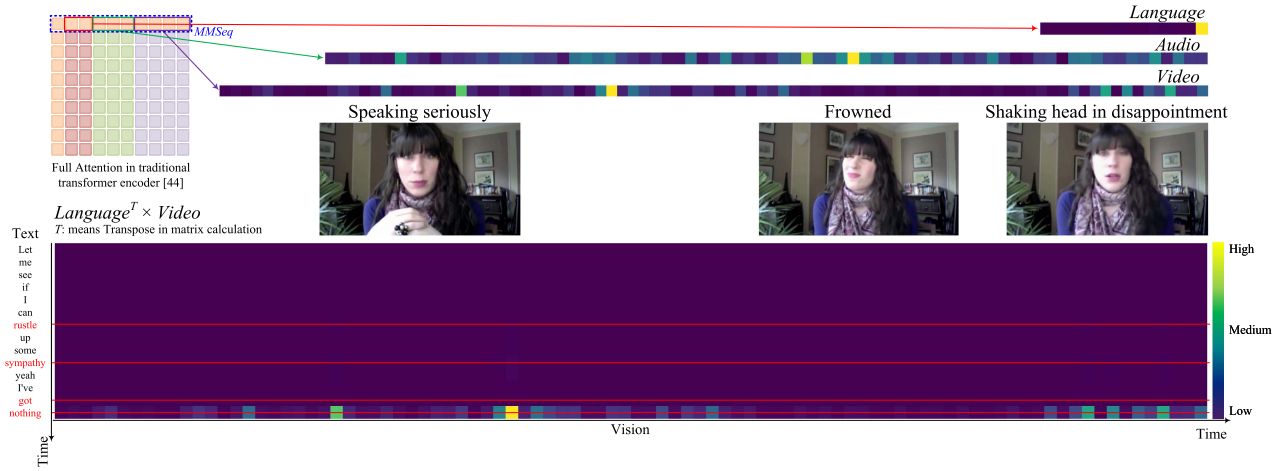
Fig. 5 demonstrates that the MUM attention mechanism of UniMF greatly improves attention on words conveying emotions other than 'nothing' in the *Language* modality. Specifically, words such as 'rustle', 'sympathy', and 'got' receive more attention, compared to the attention distribution obtained using standard full attention [53], which predominantly focuses on the word 'nothing'. Additionally, for the *Video* modality, UniMF using MUM attention can focus on both the serious speech and subsequent head-shaking motion of the heroine, whereas full attention mostly concentrates on the serious speech in the front of the video. The head-shaking motion typically delivers negative sentiment signals, thereby contributing to a more accurate sentiment prediction. We believe that the effectiveness of the MUM attention mechanism results from the exclusion of redundant cross-modal attention. This facilitates better information exchange and distinction between unimodal and multimodal information, avoiding the problems of information redundancy.

C. Visualization of Joint Representation

To further examine UniMF's robustness in the presence of different modality deficiencies, we employ the t-SNE algorithm [64] to visualize the joint multimodal representation. Fig. 6, 7, and 8 show the joint representation of UniMF, TFR-Net [23], and EMT-DLFR [39], respectively. Under various

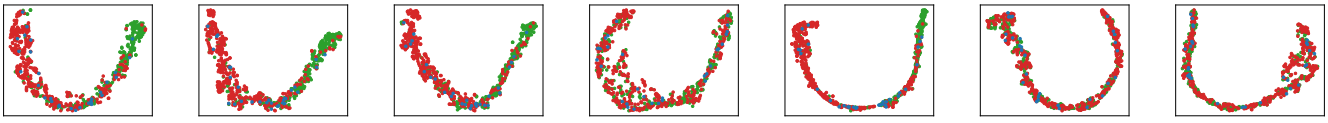


(a) MUM



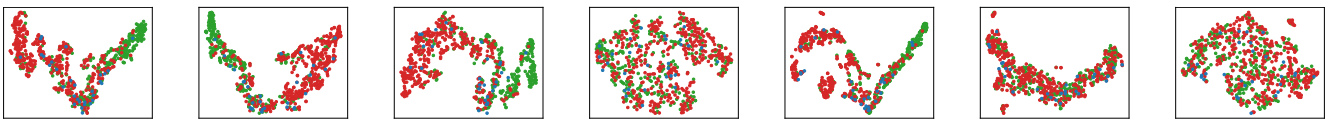
(b) Full Attention

Fig. 5. Performance of different attention mechanisms on CMU-MOSI dataset. We take out the *Language* part and *Video* part from the *MMSeq* sequence in the last layer of the transformer paying attention to. Since one video clip corresponds to one sentiment label in the CMU-MOSI dataset, we embed the special sequence *MMSeq* with the length of $T_M = 1$. Therefore, the *Language* part attended by the special sequence *MMSeq* is $1 \times T_L$ and the *Video* part is $1 \times T_V$. So, to better demonstrate the relationship between the two modalities, we do a matrix multiplication of them, thus getting a heat map of size $T_L \times T_V$.



(a) (L, A, V) (b) (L, A) (c) (L, V) (d) (A, V) (e) (L) (f) (A) (g) (V)

Fig. 6. Visualization of the joint representation of UniMF based on the CMU-MOSI dataset with different input patterns, where red indicates negative (< 0), green indicates positive (> 0), and blue indicates neutral ($= 0$).



(a) (L, A, V) (b) (L, A) (c) (L, V) (d) (A, V) (e) (L) (f) (A) (g) (V)

Fig. 7. Visualization of the joint representation of TFR-Net based on the CMU-MOSI dataset with different input patterns, where red indicates negative (< 0), green indicates positive (> 0), and blue indicates neutral ($= 0$).

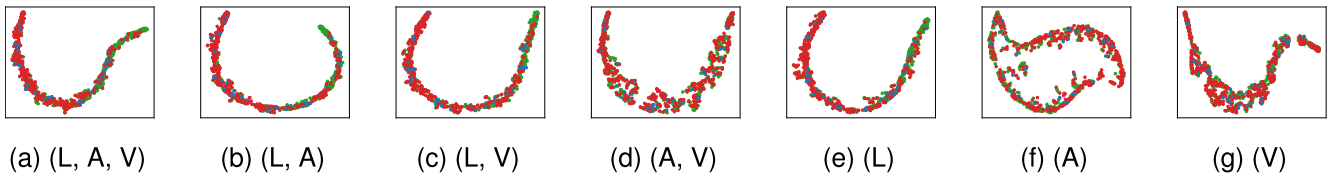


Fig. 8. Visualization of the joint representation of EMT-DLFR based on the CMU-MOSI dataset with different input patterns, where red indicates negative (< 0), green indicates positive (> 0), and blue indicates neutral ($= 0$).

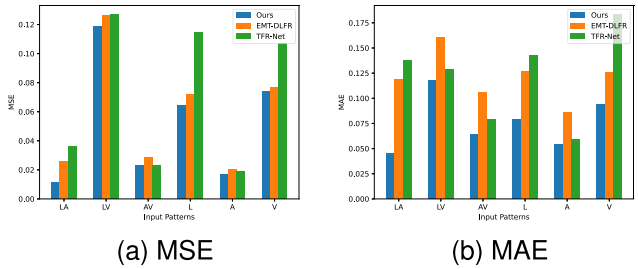


Fig. 9. Comparison of imputation performance with different input patterns on the CMU-MOSI dataset. Lower MSE and MAE indicate better imputation performance.

missing modalities, as shown in Fig. 6(b)–(g), the joint representation of UniMF remains similar to that of the full modality input case due to the joint operation of the MGT and MUT mechanisms. In contrast, the joint representation of TFR-Net under missing modalities is poorly maintained, particularly when the *Language* modality is absent, as shown in Fig. 7(d) and (g). On the other hand, although EMT-DLFR has a better joint representation distribution than TFR-Net, it fails to maintain invariance when the input pattern is (*Audio*). In summary, UniMF maintains the invariance of the joint representation distribution better than TFR-Net and EMT-DLFR when faced with modality deficiencies, thereby enhancing the final sentiment analysis performance.

D. Modality Imputation Performance

To evaluate the modality imputation performance of MGT, we calculate the mean squared error (MSE) and mean absolute error (MAE) between the generated modality and ground truth under different input patterns, and compare the results with those obtained from TFR-Net [23] and EMT-DLFR [39], as shown in Fig. 9. The results indicate that UniMF's MGT achieves smaller MSE and MAE between the generated modality and the ground truth, compared to the other imputation methods, under all cases of missing modalities. This confirms that MGT in UniMF can infer about the missing modality by observing the existing modalities' information and can achieve better modality imputation performance with fewer parameters.

VI. Conclusion

In this paper, we introduce a novel multimodal framework named UniMF, which integrates two main modules. The first is the Translation Module, which translates missing modalities via existing modality information, and its core component is the MGT, consisting of MGM that suitably combines the encoder-decoder pair in the traditional NMT transformer. The second is the Prediction Module, with the MUT as its core element, which exchanges information between unimodal and multimodal. Moreover, we introduce the *MMSeq*, a special sequence used to accumulate modality information, to enhance recognition capabilities. In conclusion, UniMF adeptly handles missing modalities and unaligned multimodal sequences concurrently, which are two major challenges in MSA, and produces competitive or SOTA results with less learnable parameters than previous methods. In the future, we will consider the process of compressing sequences, which reduces memory requirements and produces a more lightweight framework. We also aim to optimize the CUDA kernel by exploiting the sparsity of the attention matrix, accelerating the training process. Lastly, we want to address the problem of semantic information disparity between modalities and investigate better initialization methods for the `[multi]` token and *MMSeq*.

References

[1] S. Poria et al., "Context-dependent sentiment analysis in user-generated videos," in *Proc. Assoc. Comput. Linguistics*, 2017, pp. 873–883.
[2] D. Ghosal et al., "Contextual inter-modal attention for multi-modal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 3454–3466.
[3] 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.
[4] Y. Wang et al., "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7216–7223.
[5] P. P. Liang, Z. Liu, A. B. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 150–161.
[6] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Trans. Multimedia*, vol. 22, pp. 122–137, 2020.
[7] S. Mai, H. Hu, and S. Xing, "A unimodal representation learning and recurrent decomposition fusion structure for utterance-level multimodal embedding learning," *IEEE Trans. Multimedia*, vol. 24, pp. 2488–2501, 2022.
[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997 .
[9] 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. Chap. Assoc. Comput. Linguist.*, 2019, pp. 4171–4186.
[10] Z. Yang et al., "Xlnet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5754–5764.
[11] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.
[12] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.
[13] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8992–8999.

[14] K. Yang, H. Xu, and K. Gao, "Cm-bert: Cross-modal bert for text-audio sentiment analysis," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 521–528.

[15] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2276–2289, Jul.–Sep. 2023.

[16] S. Mai, Y. Zeng, and H. Hu, "Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations," *IEEE Trans. Multimedia*, vol. 25, pp. 4121–4134, 2023.

[17] H. Sun, H. Wang, J. Liu, Y.-W. Chen, and L. Lin, "Cubemlp: An MLP-based model for multimodal sentiment analysis and depression estimation," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 3722–3729.

[18] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 1642–1651.

[19] 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.

[20] H. Pham, T. Manzini, P. P. Liang, and B. Poczos, "Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis," 2018, *arXiv:1807.03915*.

[21] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proc. Web Conf.*, 2020, pp. 2514–2520.

[22] J. Tang et al., "Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proc. Assoc. Comput. Linguistics*, 2021, pp. 5301–5311.

[23] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 4400–4407.

[24] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[25] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.

[26] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.

[27] Y.-H. H. Tsai et al., "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.

[28] 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 Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2554–2562.

[29] 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 Int. Conf. Comput. Vis.*, 2021, pp. 8148–8156.

[30] J. Cheng, I. Fostiropoulos, B. Boehm, and M. Soleymani, "Multimodal phased transformer for sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2447–2458.

[31] 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. 10790–10797.

[32] 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.

[33] J. Guo, J. Tang, W. Dai, Y. Ding, and W. Kong, "Dynamically adjust word representations using unaligned multimodal information," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 3394–3402.

[34] D. Yang, H. Kuang, S. Huang, and L. Zhang, "Learning modality-specific and-agnostic representations for asynchronous multimodal language sequences," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 1708–1717.

[35] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proc. Assoc. Comput. Linguistics*, 2021, pp. 2608–2618.

[36] J. Zeng, T. Liu, and J. Zhou, "Tag-assisted multimodal sentiment analysis under uncertain missing modalities," in *Proc. Int. ACM SIGIR Conf. Res. Develop Inf. Ret.*, 2022, pp. 1545–1554.

[37] J. Zeng, J. Zhou, and T. Liu, "Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities," *IEEE Trans. Multimedia*, vol. 25, pp. 6301–6314, 2023.

[38] J. Zeng, J. Zhou, and T. Liu, "Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 2924–2934.

[39] L. Sun, Z. Lian, B. Liu, and J. Tao, "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," *IEEE Trans. Affect. Comput.*, early access, May 10, 2023, doi: 10.1109/TAFFC.2023.3274829.

[40] M. Chen et al., "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2017, pp. 163–171.

[41] S. Poria et al., "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. Assoc. Comput. Linguistics*, 2019, pp. 527–536.

[42] N. Majumder et al., "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6818–6825.

[43] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[44] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowl.-Based Syst.*, vol. 161, pp. 124–133, 2018.

[45] A. Zadeh et al., "Multi-attention recurrent network for human communication comprehension," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5642–5649.

[46] A. Zadeh et al., "Memory fusion network for multi-view sequential learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5634–5641.

[47] Z. Liu et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[48] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proc. Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[49] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–20.

[50] W. Peng, X. Hong, and G. Zhao, "Adaptive modality distillation for separable multimodal sentiment analysis," *IEEE Intell. Syst.*, vol. 36, no. 3, pp. 82–89, Mar. 2021.

[51] Z. Yuan, Y. Liu, H. Xu, and K. Gao, "Noise imitation based adversarial training for robust multimodal sentiment analysis," *IEEE Trans. Multimedia*, early access, Apr. 17, 2023, doi: 10.1109/TMM.2023.3267882.

[52] 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.*, vol. 45, no. 7, pp. 8419–8432, Jul. 2023.

[53] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[54] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[55] 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.

[56] M. K. Hasan et al., "UR-FUNNY: A multimodal language dataset for understanding humor," in *Proc. Conf. Empirical Methods Natural Lang. Process.-Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 2046–2056.

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[58] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[59] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using biLSTM-CRF and CNN," *Expert Syst. Appl.*, vol. 72, pp. 221–230, 2017.

[60] 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.

[61] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[62] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 59–66.

[63] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 2623–2631.

[64] L. Van derMaaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.