# Multi-Modality Multi-Loss Fusion Network

**Anonymous ACL submission**

## Abstract

In this work we investigate the optimal selection and fusion of features across multiple modalities and combine these in a neural network to improve emotion detection. We compare different fusion methods and examine the impact of multi-loss training within the multi-modality fusion network, identifying surprising findings relating to subnet performance. Our best model achieves state-of-the-art performance for three datasets (CMU-MOSI, CMU-MOSEI and CH-SIMS), and outperforms the other methods in most metrics. We have found that training on multimodal features improves single modality testing and designing fusion methods based on dataset annotation schema enhances model performance. These results suggest a roadmap toward an optimized feature selection and fusion approach for enhancing emotion detection in neural networks.

## 1 Introduction

The multimodal affective computing field has seen significant advances in feature extraction and multimodal fusion methodologies in recent years. By combining audio, text and visual signals, these models offer a more comprehensive, nuanced understanding of human emotions. However, there are still limitations: hand-crafted feature extraction algorithms often lack flexibility and generalization across diverse tasks. To overcome these limitations, recent studies have proposed fully end-to-end models that optimize both feature extraction and learning processes jointly (Dai et al., 2021). Our work extracts feature representations from pre-trained models for different modalities and combines them in an end-to-end manner, which provides a comprehensive and adaptable solution for multimodal affective feature computation. In the context of multimodal fusion, the challenge also lies in effectively fusing diverse signals, including natural language, facial gestures, and acoustic behaviors. Methods like Tensor Fusion Network (TFN)(Zadeh et al., 2017) have been proposed to model intra-modality and inter-modality interactions. More recently, transformer encoder structures such as MULT(Tsai et al., 2019) with cross-modal attention have gained popularity for integrating multimodal data. In this paper, we propose a novel fusion network structure that integrates cross-modal attention and self-attention, with additional feed-forward layers to refine the representations. We have also experimented with variations of the proposed fusion network.

Despite the recent advancements, there is still room for improvement in multimodal feature extraction and fusion. In this study, we present a series of experiments that focus on feature selection, fusion network performance comparison, and multi-loss training analysis using audio and text data from three datasets: CMU-MOSI, CMU-MOSEI, and CH-SIMS. We compare different methods for extracting audio features as well as different fusion network methods to combine audio and text signals to identify the best-performing procedures. We find that the addition of audio signals consistently improves performance and also that our transformer fusion network further enhances results for most metrics and achieves state-of-the-art results across all datasets, indicating its efficacy in enhancing cross-modality modeling and its potential for multimodal emotion detection. From multi-loss training, we also observe that 1) using distinct labels for each modality in multi-loss training significantly benefits the models' performance, and 2) training on multimodal features improves not only the overall model performance but also the model's accuracy on the single-modality subnet. These novel findings have advanced our understanding of multimodal sentiment analysis and hold promise for further research and optimization in this field.

* These authors contributed equally to this work.

## 2 Related Work

Existing research on multimodal affective computing often employs hand-crafted algorithms to perform initial feature representation extraction and retrieve some fixed representations for each modality (Shenoy and Sardana, 2020). (Delbrouck et al., 2020) However, for these, the extracted features are static and lack the flexibility to be further fine-tuned for different target tasks; also, the manual determination of feature extraction algorithms can lead to sub-optimal performance due to constraints in generalization across diverse tasks (Dai et al., 2021). To address these issues, recent studies have proposed fully end-to-end models, effectively bridging the gap between feature extraction and learning processes (Dai et al., 2021) (Wang et al., 2020). Our research also emphasizes an end-to-end structure that optimizes both phases jointly, presenting a comprehensive and adaptable solution for multimodal affective feature computation.

Lexical features, owing to pre-training on expansive corpora through Transformer-based models, often outperform other modalities. Some recent work aims to improve model performance by incorporating speech information inside the text model such as SPECTRA (Yu et al., 2023), by pretraining a speech-text transformer model to capture the speech-text alignment effectively. A similar innovative method is the Transformer-Based Speech-Prefixed Language Model (TEASEL) (Arjmand et al., 2021), which incorporates speech as a dynamic prefix along with the textual.

Many studies have explored multimodal human language time-series data, which typically includes a mixture of natural language, facial gestures, and acoustic behaviors. However, fusing these into a unified representation presents a significant challenge due to the variable sampling rates across modalities and the difficulty in determining intra-modality dependencies. Various methods have been proposed to model the interaction across modalities, such as the Tensor Fusion Network (Zadeh et al., 2017), which utilizes the Cartesian product of different modalities to model both intra-modality and inter-modality interactions. More recent work has shifted toward employing transformer encoder structures to integrate these signals via cross-modality attention. The MULT model (Tsai et al., 2019) has pioneered this approach, introducing directional pairwise cross-modal attention. This method allows for interaction between multimodal sequences across distinct time steps and inherently adapts streams from one modality to another. Further research has also leveraged this concept of cross-modality attention (Goncalves and Busso, 2022) (Paraskevopoulos et al., 2022), yielding valuable insights into how multimodal data can be processed more effectively. We enhance this approach by employing a self-attention encoder and a feed-forward network to further optimize the multimodal representation after one modality is projected into another using the cross-modality attention module, thus enriching our ability to process and understand multimodal data.

## 3 Methodology

The methodology for emotion detection in our study involves two primary components: the feature network and the fusion network. Each of these has its own unique mechanisms and contributes towards the overall functioning of our proposed Multi-Modality Multi-Loss Fusion Network (MMML) as illustrated in Figure 1.

### 3.1 Feature Network

The Feature Network employs two different pre-trained models for text and audio processing. The text subnet leverages RoBERTa, chosen for its significantly superior performance on various downstream tasks. The audio subnet employs different models for different languages: HuBERT for Mandarin and Data2Vec for English. This ensures the optimized extraction of features from the given modalities, setting a solid foundation for the subsequent fusion process.

### 3.2 Fusion Network

The Fusion Network is the heart of the MMML, where the information from multiple modalities is combined. This network is divided into three smaller components as shown in the yellow portion of Figure 1. First, there is a Cross-Attention Encoder, which adopts a mechanism similar to the self-attention encoder but which employs a query from one modality and uses keys and values generated from another modality. This cross-modal interaction aims to capture the inter-dependencies between different modalities, contributing to a more holistic understanding of the data. This encoder is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$


Figure 1: Our Model Structure

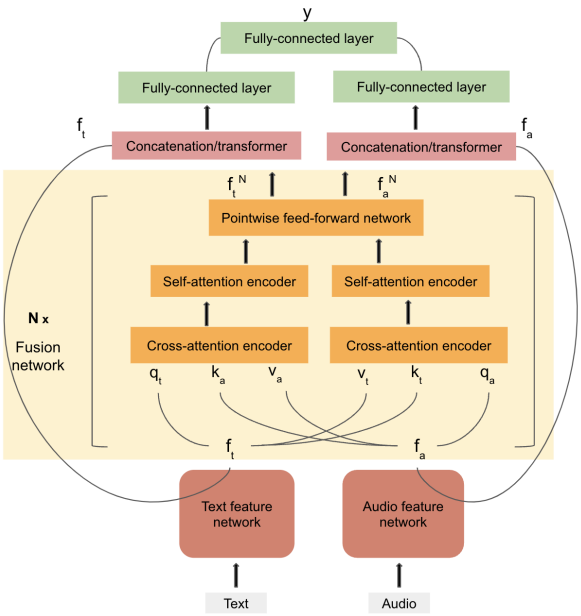


Figure 2: Model Variations

where Q is the matrix of queries, K is the matrix of keys, V is the matrix of values, and $d_k$ is the dimension of the keys.

In cross-modality attention, we denote queries as $Q_{m1}$ (from modality 1) and the keys and values as $K_{m2}$ and $V_{m2}$ (from modality 2). $f_{m1}$ is the feature from modality 1 and $f_{m2}$ is the feature from modality 2. The formula is:

$$\text{Attention}(Q_{m1}, K_{m2}, V_{m2}) = \text{softmax}\left(\frac{Q_{m1}K_{m2}^T}{\sqrt{d_k}}\right)V_{m2}$$

where:

$$Q_{m1} = W_q \cdot f_{m1}$$
$$K_{m2} = W_k \cdot f_{m2}$$
$$V_{m2} = W_v \cdot f_{m2}$$

Our proposed network also includes additional *Self-Attention Encoders* which use a traditional self-attention mechanism, as found in the original transformer models and is designed to find the correlation within a single modality, thereby capturing the intra-modal dynamics of the data. In our model, the self-attention module serves to model the connections across time steps of the new feature representation after passing through the cross-modality encoder.

Finally, it includes a *Pointwise Feed-Forward Network* which applies fully connected feed-forward networks and ReLU activation functions to each individual position, further refining the encoded feature representations. Through combining these methodologies, we aim to optimize the multi-modal feature extraction and fusion process, enhancing the MMML's performance in emotion detection tasks.

The operationalization of our methodology involves processing the raw text and audio data through their respective pre-trained models. The final hidden states obtained from both the text and audio subnets are employed in two experimental ways: 1) *Direct Concatenation*, in which the mean of the hidden states from different modalities is computed and utilized as the feature set. These mean features are then directly concatenated to represent the combined information from all modalities; 2) *Fusion Network*, which incorporates additional layers of the fusion network before concatenation, as illustrated in the yellow portion of Figure 1. The process begins with the introduction of a CLS token pre-pended to the hidden states for each modality which then serve as markers for the final modality representation.

A critical piece of this fusion approach is the role of the cross-attention encoder, which initiates the fusion process by generating a query from the text modality and inquiring keys from the audio modality. It seeks to identify which values from the audio modality relate to each text segment and to then provide each text segment with a weighted average of audio hidden states. This interaction is critical because the cross-attention encoder essentially projects the hidden states from one modality into the space of another modality. After this inter-modal projection, a self-attention layer is added to

find the correlation within the newly formed space of hidden states. This layer allows for the capture of intra-modal dynamics within this new mixed space.

After completing these stages of interaction and transformation, the final output values from the CLS tokens are used as the final modality representation. The concatenated representation is then passed through a series of three fully connected layers to generate the final prediction.

The fusion network approach aims to enhance the integration of multi-modal information and provide more in-depth insight into the correlations and interdependencies between different modalities.

### 3.3 Multi-Loss Training

In order to leverage *multi-loss training*, we modified the architecture of our fusion network to incorporate an additional fully-connected layer at the termination of each feature network, as illustrated in the green portion of Figure 1. This modification enables two additional outputs from individual modalities, in addition to the combined feature output. This design facilitates the application of three distinct loss functions during training, each corresponding to one of the outputs.

The rationale behind implementing additional losses for each individual modality is to bolster the respective feature networks' comprehension and processing of their respective signals. Given that each feature network perceives and handles signals distinctively, akin to how humans discern emotions through different sensory signals, the multi-task loss serves a dual purpose: First, it encourages each feature network to refine its method of processing its specific modality, akin to honing the 'sense' associated with that modality; Second, it trains the fusion network to effectively combine the distinct signals relayed by the feature networks, as guided by the loss from the combined modality.

Through this multi-loss training approach, we create a model that efficiently mirrors human-like multi-modal emotion perception, each modality working independently and collaboratively to understand the comprehensive emotional context.

### 3.4 Variations of the Fusion Network

In our investigation of the fusion network, we developed two variants designed to mitigate potential loss of original signals during the cross-modal projection process. Because the cross-attention mechanism projects one modality into another, some original signals might be obscured or lost. Therefore, these variations aim to combine the original signal with the projected signal, thereby enhancing the ability of the network to learn from both signals simultaneously:

The first is *Concatenation Variation* which concatenates the original feature with the fused feature. This fusion of original and projected information within each modality aims to maintain the integrity of the original signals, while also integrating the enriched cross-modal information. The combined features then go through a linear layer and are subsequently concatenated with features from other modalities.

The second is the *Transformer Variation*. This variation merges the original hidden states and the fused hidden states along the feature dimension. Transformer encoders further process these combined hidden states. Finally, the mean of the hidden states serves as the feature for that modality, and these features are concatenated. The transformational capabilities of the transformer network are leveraged to refine and combine the feature sets, which are then used for the final prediction.

By incorporating both original and projected signals, these variations offer a more comprehensive feature set for the final layers of the model, enhancing the model's ability to accurately detect and interpret emotions from multimodal inputs. These variations provide a comprehensive exploration of feature fusion strategies.

## 4 Experiments

### 4.1 Experimental Setup

We use three primary datasets, each characterized by its unique properties and content, to test the performance of the Multi-Modality Multi-Loss Fusion Network on emotion detection.

The *CMU-Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSI)*(Zadeh et al., 2016): This dataset, developed in English, comprises audio, text, and video modalities compiled from 2199 annotated video segments collected from YouTube monologue movie reviews. It offers a focused approach to studying emotion detection within the context of film critique.

The *CMU-Multimodal Sentiment Analysis (CMU-MOSEI)*(Bagher Zadeh et al., 2018) is an extension of CMU-MOSI, including the same modalities of audio, text, and video from YouTube videos, but it has a broader scope, covering a wider range of

| | Has0_ACC$_2$ | Has0_F1 | Non0_ACC$_2$ | Non0_F1 | ACC$_5$ | ACC$_7$ | MAE | CORR |
|---|---|---|---|---|---|---|---|---|
| TEASEL | 84.79 | 84.72 | 87.5 | 85 | - | 47.52 | 64.4 | 83.6 |
| SPECTRA | - | - | 87.5 | - | - | - | - | - |
| UniMSE | 85.85 | 85.83 | 86.9 | 86.42 | - | **48.68** | 69.1 | 80.9 |
| **MMML (Ours)** | **85.91** | **85.85** | **88.16** | **88.15** | **56.08** | 48.25 | **64.29** | **83.8** |

(a) CMU-MOSI

| | Has0_ACC$_2$ | Has0_F1 | Non0_ACC$_2$ | Non0_F1 | ACC$_5$ | ACC$_7$ | MAE | CORR |
|---|---|---|---|---|---|---|---|---|
| SPECTRA | - | - | 87.34 | - | - | - | - | - |
| UniMSE | 85.86 | 85.79 | **87.5** | **87.46** | - | 54.39 | 52.3 | 77.3 |
| **MMML (Ours)** | **86.32** | **86.23** | 86.73 | 86.49 | **57.32** | **54.95** | **51.74** | **79.08** |

(b) CMU-MOSEI

| | ACC$_2$ | ACC$_3$ | ACC$_5$ | F1 | MAE | CORR |
|---|---|---|---|---|---|---|
| EMT | 80.1 | 67.4 | 43.5 | 80.1 | 39.6 | 62.3 |
| **MMML(ours)** | **82.93** | **69.37** | **49.38** | **82.9** | **33.2** | **73.26** |

(c) CH-SIMS

Table 1: **Comparison with SOTA**: All three datasets achieve state-of-the-art performance. All experimental results presented are averages derived from three separate runs.

topics, and is more substantial in size, with 23,453 annotated video segments.

The *Chinese Multimodal Sentiment Analysis Dataset (CH-SIMS)*(Yu et al., 2020), a Mandarin language dataset incorporating the same modalities: audio, text, and video, collected from 2281 annotated video segments. It comprises data from TV shows and movies, making it culturally distinct and diverse, and includes multiple labels for the same utterance based on different modalities, which adds an extra layer of complexity and richness to the data.

These datasets provide a broad and multicultural perspective on emotion detection, allowing for a thorough evaluation and comparative analysis of the MMML's performance across diverse data landscapes.

Our MMML model was evaluated using metrics consistent with existing research against existing benchmarks, which enables comprehensive evaluation of our model's performance across diverse sentiment analysis dimensions (detailed descriptions in Appendix A.2). Additional sets of ablation experiments on different components of the model were conducted for analysis, to interpret and explain the model performance.

### 4.2 Results

Our overall results are shown in Table 1. When compared with contemporary state-of-the-art models, our method emerges as a robust performer, offering superior outcomes for both CMU-MOSI and CMU-MOSEI. Among recent models, UniMSE (Hu et al., 2022) has delivered good results on the English datasets. Nonetheless, our MMML model surpasses UniMSE in most of the evaluation metrics, reinforcing the effectiveness of our approach. Intriguingly, there is little research on CH-SIMS for emotion detection tasks. However, one state-of-the-art model is the Efficient Multimodal Transformer (EMT) (Sun et al., 2023), which has demonstrated a high degree of performance over existing methods. Our MMML model significantly outperforms EMT across all metrics, further underscoring the potential of our fusion network. These results not only validate our multimodal fusion network but also affirm the robustness of our chosen methodology for emotion detection tasks. Our impressive performance on all three datasets, CMU-MOSI, CMU-MOSEI, and CH-SIMS, verifies the versatility and adaptability of our MMML model, emphasizing its value in advancing the field of emotion detection.

### 4.3 Audio Feature Selection

To incorporate the best speech information into our model, the initial stage of our experimentation process involved comparing the performance on audio features for sentiment analysis from two datasets, CMU-MOSI (English) and CH-SIMS (Mandarin), using openSMILE and Mel spectrograms, each with customized parameters for optimal feature extraction to compare with features from a pre-trained audio model. Implementation details are in Appendix A.2.

| Feature name | $\mathbf{ACC_2}$ |
|---|---|
| openSMILE | 0.6696 |
| Mel Spectrogram | 0.6805 |
| Fine-tuned HuBert(CH) | **0.7465** |

(a) CH-SIMS

| Feature name | $\mathbf{ACC_2}$ |
|---|---|
| openSMILE | 0.4606 |
| Mel Spectrogram | 0.4519 |
| Fine-tuned Data2vec(EN) | **0.7099** |

(b) CMU-MOSI

Table 2: **Audio Feature Selection Results**: Fine-tuning a pre-trained audio model works significantly better than using other audio features.

Upon evaluation of the different audio feature extraction methods shown in Table 2, we found that use of a pre-trained model for raw audio yielded higher accuracy rates: accuracy rates of approximately 71% and 75% were achieved for CMU-MOSI and CH-SIMS, respectively. This outperformed the other two techniques (openSMILE and Mel spectrograms) by a significant margin. Interestingly, openSMILE and Mel spectrograms displayed comparable performance on CH-SIMS. However, their performance on CMU-MOSI was notably subpar. We hypothesize that CH-SIMS, comprising audio from TV shows and movies, presents a more straightforward task for audio emotion classification.

This analysis highlights the effectiveness of using pre-trained models for raw audio in achieving superior emotion classification accuracy. It also underscores the need to consider the characteristics and source of audio data in applying different feature extraction techniques.

### 4.4 Comparison of Simple Concatenation and Fusion Network

To prove the superiority of our proposed fusion network, we compared it against concatenation. Upon analysis of our results, as shown in Table 3, we observed that the introduction of the transformer fusion network yielded improvements in performance in most metrics for CMU-MOSEI and CH-SIMS, and half of the metrics for CMU-MOSI. These results underscore the effectiveness of our transformer fusion network in enhancing cross-modality modeling and suggest its potential as a powerful tool for multi-modal emotion detection.

Beyond these observations, it is imperative to highlight that both methods which combined audio and text signals outperformed methods utilizing only text signals in almost all metrics across the three datasets. A noteworthy increase in performance was recorded on the CH-SIMS dataset upon the addition of audio signals, while the two English datasets, CMU-MOSI and CMU-MOSEI, exhibited smaller improvements. The substantial improvement observed in CH-SIMS can be attributed to two factors. First, CH-SIMS assigns unique labels to audio and text, thereby facilitating the network's ability to learn distinct signals from each modality. Second, the source for CH-SIMS is TV show and movie videos, which typically display easily-interpretable emotions. This characteristic probably contributes to the effectiveness of combining audio and text signals for emotion detection.

### 4.5 Multi-loss Training Experiments

To investigate the effectiveness of multi-loss training, we performed comparative experiments on two different datasets: CMU-MOSEI and CH-SIMS. CMU-MOSEI provides a single target for each utterance, whereas CH-SIMS offers different labels for each modality in addition to the combined modalities. We easily adapted multi-loss training to CH-SIMS, given its distinct labels for each modality. For CMU-MOSEI, we duplicated the single target across different losses to enable multi-task training.

The results, as show in Table 4, were striking: while multi-loss and single-loss training performed similarly on CMU-MOSEI, multi-loss training significantly boosted performance on CH-SIMS. This underscores the value of unique labels for each modality when employing multi-task training. The improved performance on CH-SIMS can be attributed to the distinct nature of the signals processed by each feature network. Since audio includes acoustic signals that are not present in the text, it is common for them to have different sentiment. Having distinct labels assists each network in learning better how to process its unique signal.

Surprisingly, as shown in Table 5, the multi-loss training also contributed to an enhanced performance of the text subnet when compared to training with only the text. The additional audio signal appears to support the performance improvement. This suggests that even when the goal is to use only the text input for inference, multi-loss training

| | $\mathbf{Has0\_ACC_2}$ | Has0_F1 | $\mathbf{Non0\_ACC_2}$ | Non0_F1 | $\mathbf{ACC_5}$ | $\mathbf{ACC_7}$ | MAE | CORR |
|---|---|---|---|---|---|---|---|---|
| **text-only** | 84.79 | 84.72 | 87.29 | 87.29 | 56.41 | 48.68 | 64.96 | 83.61 |
| **concatenation** | 85.77 | 85.74 | 87.6 | 87.62 | **56.51** | **48.79** | **64.27** | **84.06** |
| **+ fusion network** | **85.91** | **85.85** | **88.16** | **88.15** | 56.08 | 48.25 | 64.29 | 83.8 |

(a) CMU-MOSI

| | $\mathbf{Has0\_ACC_2}$ | Has0_F1 | $\mathbf{Non0\_ACC_2}$ | Non0_F1 | $\mathbf{ACC_5}$ | $\mathbf{ACC_7}$ | MAE | CORR |
|---|---|---|---|---|---|---|---|---|
| **text-only** | 84.81 | 84.95 | 86.34 | 86.19 | 54.99 | 52.7 | 53.31 | 78.6 |
| **concatenation** | 84.77 | 84.9 | **86.82** | **86.65** | 55.99 | 53.94 | 51.63 | **79.81** |
| **+ fusion network** | **86.32** | **86.23** | 86.73 | 86.49 | **57.32** | **54.95** | **51.54** | 79.08 |

(b) CMU-MOSEI

| | $\mathbf{ACC_2}$ | $\mathbf{ACC_3}$ | $\mathbf{ACC_5}$ | F1 | MAE | CORR |
|---|---|---|---|---|---|---|
| **text-only** | 79.21 | 65.06 | 42.02 | 79.14 | 42.65 | 59.4 |
| **concatenation** | 81.91 | **70.68** | 47.12 | 82.1 | 34.96 | 72.37 |
| **+ fusion network** | **82.93** | 69.37 | **49.38** | **82.9** | **33.2** | **73.26** |

(c) CH-SIMS

Table 3: **Concatenation vs. Transformer Fusion**: Integration of audio signals enhances performance across almost all metrics, with more pronounced impact on CH-SIMS. Implementing the Fusion Network augments performance slightly in most metrics. All experimental results presented are averages derived from three separate runs.

can be beneficial. The text subnet, after training with the multi-modal model, can be extracted and used independently, offering superior performance compared to when it is trained alone.

Interestingly, this improvement was not observed in the audio subnet, potentially due to the stronger signal from the text subnet (reflected by a 10% higher accuracy when trained alone) which made it easier to train, and thus the network might have focused on reducing its loss.

In summary, the benefits of multi-loss training are threefold. First, it substantially boosts the performance of the entire network when distinct labels for different modalities are available. Second, loss from other modalities enhances the performance of the text subnet, indicating that we can utilize other modalities in training even when the text subnet is the only required component for inference. Third, it is capable of handling missing modalities, enabling outputs when only text or audio inputs are available. These findings shed light on the potential of multi-loss training in the context of multi-modality fusion networks, opening avenues for further research and optimization.

### 4.6 Result for Fusion Network Variations

To understand the effect of restoring original signals, we conducted a comparative analysis of proposed fusion network variations, which reveals a relatively consistent performance across all variations. As shown by the results presented in Table 6, the three methods demonstrate similar performance across all metrics for both CMU-MOSEI and CH-SIMS. Surprisingly, reincorporating the original signal into the fused signal did not lead to any significant improvement in performance. In essence, while similar performance across different fusion network variations was unanticipated, it paves the way for a deeper understanding of the interactions within the fusion network and the role of original signals in such approaches.

## 5 Conclusion

In conclusion, this study has provided novel, important findings for multi-modal sentiment analysis that should benefit future researchers in the designing of sentiment analysis and other models, presenting a SOTA model. First, the use of pre-trained models for raw audio yielded superior results, highlighting their effectiveness in feature extraction. Second, combining audio and text signals consistently outperformed using text signals alone, with the transformer fusion network showing promise in enhancing cross-modality modeling. Third, multi-loss training proved beneficial, particularly with unique labels for each modality. Last, achieving state-of-the-art results on three emotion detection datasets underscores the effectiveness of our approach. Still, the performance of fusion network variations did remain consistent, prompting further investigation.

| | Has0_ACC$_2$ | Has0_F1 | Non0_ACC$_2$ | Non0_F1 | ACC$_5$ | ACC$_7$ | MAE | CORR |
|---|---|---|---|---|---|---|---|---|
| **single-loss** | **85.22** | **85.39** | **87.02** | **86.91** | 55.95 | 53.85 | 51.96 | 79.68 |
| **multi-loss** | 84.77 | 84.9 | 86.82 | 86.65 | **55.99** | **53.94** | **51.63** | **79.81** |

(a) CMU-MOSEI

| | ACC$_2$ | ACC$_3$ | ACC$_5$ | F1 | MAE | CORR |
|---|---|---|---|---|---|---|
| **Single-loss** | 78.34 | 67.18 | 46.83 | 78.59 | 39.09 | 62.69 |
| **multi-loss** | **81.91** | **70.68** | **47.12** | **82.1** | **34.96** | **72.37** |

(b) CH-SIMS

Table 4: **Single-Loss Training vs. Multi-Loss Training**: While multi-loss training does not yield performance improvement when identical labels are used for different losses, as in the case of CMU-MOSEI, it does contribute significantly to performance enhancement when unique labels are assigned to each modality, as observed with CH-SIMS.

| | Has0_ACC$_2$ | Has0_F1 | Non0_ACC$_2$ | Non0_F1 | ACC$_5$ | ACC$_7$ | MAE | CORR |
|---|---|---|---|---|---|---|---|---|
| **text-loss only** | 84.79 | 84.72 | 87.29 | 87.29 | **56.41** | **48.68** | 64.96 | 83.61 |
| **multi-loss** | **85.62** | **85.56** | **87.91** | **87.9** | 55.01 | 47.42 | **64.76** | **83.79** |

(a) CMU-MOSI

| | Has0_ACC$_2$ | Has0_F1 | Non0_ACC$_2$ | Non0_F1 | ACC$_5$ | ACC$_7$ | MAE | CORR |
|---|---|---|---|---|---|---|---|---|
| **text-loss only** | **84.81** | **84.95** | 86.34 | 86.19 | 54.99 | 52.97 | 53.31 | 78.6 |
| **multi-loss** | 84.36 | 84.62 | **86.85** | **86.76** | **56.06** | **53.61** | **52.35** | **79.49** |

(b) CMU-MOSEI

| | ACC$_2$ | ACC$_3$ | ACC$_5$ | F1 | MAE | CORR |
|---|---|---|---|---|---|---|
| **text-loss only** | 79.21 | 65.06 | 42.02 | 79.14 | 42.65 | 59.4 |
| **multi-loss** | **83.15** | **72.14** | **48.21** | **83.74** | **28.58** | **78.72** |

(c) CH-SIMS

Table 5: **Impact of Multi-Loss on Text Subnet**: Utilizing audio-related losses can enhance performance of the text subnet, even when identical labels are employed, as is the case with CMU-MOSEI. Remarkably, using specific labels for different modalities results in a substantial performance boost in the text subnet, as evidenced by the results from CH-SIMS.

| | Has0_ACC$_2$ | Has0_F1 | Non0_ACC$_2$ | Non0_F1 | ACC$_5$ | ACC$_7$ | MAE | CORR |
|---|---|---|---|---|---|---|---|---|
| **Fused Features Only** | **86.32** | **86.23** | 86.73 | 86.49 | **57.32** | 54.95 | **51.54** | 79.08 |
| **Concatenation** | 84.96 | 85.09 | **86.78** | **86.61** | 56.86 | **57.78** | 51.88 | **79.09** |
| **Transformer** | 86.11 | 86.08 | 86.7 | 86.46 | 57.01 | 54.31 | 51.97 | 78.96 |

(a) CMU-MOSEI

| | ACC$_2$ | ACC$_3$ | ACC$_5$ | F1 | MAE | CORR |
|---|---|---|---|---|---|---|
| **Fused Features Only** | **82.93** | 69.37 | 49.38 | **82.9** | 33.2 | **73.26** |
| **Concatenation** | 82.42 | 69.44 | 49.82 | 82.38 | 33.6 | 72.87 |
| **Transformer** | 82.42 | **69.95** | **49.89** | 82.52 | **33.12** | 72.61 |

(b) CH-SIMS

Table 6: **Comparative Performance of Model Variations**: The *Fused Features Only* model employs only the features following the fusion network, while the *Concatenation* model merges the original signal with the fused signal. The *Transformer* model uses a transformer to combine these two signals. Across all metrics for both CMU-MOSEI and CH-SIMS, these three methods exhibit similar performance.

## 6 limitations

One limitation of this paper is that the proposed model is studied under two language, English and Mandarin. Another is that the coverage of domains is limited to the design of the datasets we choose to use, which is from YouTube Videos and TV shows. Hence, it's likely that a portion of the data is acted rather than naturally occurring in real life, and acted emotions may be expressed differently than natually occurring emotions.

Another limitation is that among the 3 public dataset we used, which are all collected from YouTube and TV shows, not all have detailed descriptions about anonymization of the persons appeared in the dataset. However, we did not modify the dataset, since the datasets are widely used and we would like to create coherent and comparable results with previous work.

As for potential risk of misuse, since the paper is focus on more fundamental sinde of the research, it's possible that the model might not perform well if deployed in other scenarios without additional fine-tuning and training, because the model is trained on public dataset collected from TV shows and YouTube. Misuse of directly deploying the model into real-life applications create risks as the prediction will not always be accurate.

## References


Mehdi Arjmand, Mohammad Javad Dousti, and Hadi Moradi. 2021. Teasel: A transformer-based speech-prefixed language model.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung. 2021. Multimodal end-to-end sparse model for emotion recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5305–5316, Online. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 1–7, Seattle, USA. Association for Computational Linguistics.

Lucas Goncalves and Carlos Busso. 2022. Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features. *IEEE Transactions on Affective Computing*, 13(4):2156–2170.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Georgios Paraskevopoulos, Efthymios Georgiou, and Alexandros Potamianos. 2022. Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis.

Aman Shenoy and Ashish Sardana. 2020. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*. Association for Computational Linguistics.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–17.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences.

Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020*, WWW '20, page 2514–2520, New York, NY, USA. Association for Computing Machinery.

Tianshu Yu, Haoyu Gao, Ting-En Lin, Min Yang, Yuchuan Wu, Wentao Ma, Chao Wang, Fei Huang, and Yongbin Li. 2023. Speech-text dialog pre-training for spoken dialog understanding with explicit cross-modal alignment.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos.

# A Appendix

## A.1 Training Details

The training process employed a learning rate of 1e-5, batch size of 16, and the AdamW optimizer. L2 loss was used to optimize the model during the training process. The validation set loss and accuracy were monitored to ensure the model was not overfitting to the training data. An early stopping mechanism with patience of 8 epochs was employed to ensure the generalizability of the model. The entire procedure was conducted on a single RTX 4090 GPU. For the audio pre-trained model, the Convolutional Neural Network (CNN) portion, used for feature extraction, was frozen. The impact of different learning rates for various parts of the network was explored, but no significant differences were observed. Moreover, We found that using 5 layers of fusion network achieves the best results. All the results presented in the tables are averaged over three independent runs.

## A.2 Audio Feature Extraction and Modeling Details

For openSMILE, we manipulated the frame size and step, setting them to 0.06 seconds and 0.02 seconds respectively. For Mel spectrograms, the number of Mel filterbanks was set to 128, while the window size and step were adjusted to 0.06 seconds and 0.02 seconds respectively. These configurations were chosen to enhance the precision of audio feature extraction without sacrificing computational efficiency.

Following the feature extraction phase, these features were used to construct models with varying architectures: Transformer models, incorporating between 2 and 4 encoder layers complemented with positional encoding, were employed to process the openSMILE features. A feed-forward layer was subsequently added to process feature embeddings in the CLS token of the final transformer encoder. For processing Mel spectrogram features, we leveraged convolutional neural network (CNN) models, including a custom 8-layer CNN model and modified versions of ResNet-18 and ResNet-32. The choice of these CNN architectures was driven by their known effectiveness in handling image-like data structures such as spectrograms.

## A.3 Metrics

Our MMML model was evaluated using metrics consistent with existing research.

For CMU-MOSI and CMU-MOSEI, we used:

- $\mathbf{Has0\_ACC_2}$, $\mathbf{Has0\_F1}$, including zero sentiment scores as positive;
- $\mathbf{Non0\_ACC_2}$, $\mathbf{Non0\_F1}$, ignoring zero sentiment scores;
- $\mathbf{ACC_5}$, $\mathbf{ACC_7}$, represent 5-class and 7-class accuracies respectively;
- $\mathbf{MAE}$, Mean Absolute Error;
- $\mathbf{CORR}$, assesses correlation between predicted and actual scores.

For CH-SIMS, we utilized:

- $\mathbf{ACC_2}$, $\mathbf{ACC_3}$, $\mathbf{ACC_5}$, represent 2-class, 3-class, and 5-class accuracies respectively;
- $\mathbf{F1}$, balances precision and recall;
- $\mathbf{MAE}$, mean absolute error;
- $\mathbf{CORR}$, assesses correlation between predicted and actual scores.

These metrics enable comprehensive evaluation of our model's performance across diverse sentiment analysis dimensions.