A Mixture-of-Experts Model for Multimodal Emotion Recognition in Conversations

Anonymous ACL submission

Abstract

Emotion Recognition in Conversations (ERC) requires modeling the temporal context of multi-turn dialogues and the complementary information across modalities. We propose **Mi**xture of **S**peech-**T**ext **E**xperts for **R**ecognition of **E**motions (MiSTER-E), a modular Mixture-of-Experts (MoE) framework that decouples modality-specific context modeling from multimodal integration. MiSTER-E incorporates LLM-based representations for speech and text, uses a convolutional-recurrent layer for context modeling, and integrates unimodal and cross-modal information through a gating mechanism. We introduce a supervised contrastive loss between aligned speech and text representations and a KL-divergence-based regularization to encourage agreement across expert predictions. Notably, our method does not rely on speaker identity during training or inference. Experiments on two benchmark datasets—IEMOCAP and MELD—show that our proposal achieves 70.9% and 69.5% weighted F1-scores respectively, outperforming prior speech-text ERC models. We also provide various ablations to highlight the contributions made in the proposed approach.

1 Introduction

Emotion Recognition in Conversation (ERC) seeks to infer emotional states from multi-turn, multimodal interactions. As a core task for building socially aware AI, ERC enables a range of applications including dialogue systems (Pantic et al., 2005), social media analysis (Gaind et al., 2019), and mental health monitoring (Ghosh et al., 2019). Emotions are conveyed through diverse modalities—textual content, vocal prosody, and visual cues—and evolve across conversational contexts. This layered complexity of multimodal expression of emotions makes ERC challenging.

Prior work has advanced ERC through contextual modeling (Hazarika et al., 2018; Majumder et al., 2019), speaker-aware representations (Hu et al., 2021; Shen et al., 2025), and fusion strategies ranging from early concatenation (Han et al., 2021) to attention-based and tensor methods (Zadeh et al., 2017; Hazarika et al., 2020; Dutta and Ganapathy, 2022). Most existing approaches conflate the two distinct modeling challenges: temporal context modeling and cross-modal fusion. This design choice, especially under the small size of ERC datasets, risks overfitting. This raises a core research question: *Can architectural modularity, which disentangles context modeling from modality fusion, enable improved ERC?*

In this work, we explore this question by designing an ERC framework that separates modality-specific contextual modeling from multimodal fusion. Our architecture, **Mi**xture of **S**peech-**T**ext **E**xperts for **R**ecognition of **E**motions (MiSTER-E), is structured not to optimize performance alone, but to segregate the contributions of each modality and their interactions via a Mixture-of-Experts (MoE) design. Our approach trains large language models for enhanced utterance level modeling, followed by convolutional-BiGRU networks for modeling conversational dynamics of each modality. A third branch performs multimodal modeling using cross-attention and self-attention layers. These three expert branches — speech, text, and speech-text — are then integrated through a gating mechanism that adaptively weighs each expert's prediction. To encourage cross-modal alignment, we introduce a modality-aware contrastive loss that aligns text and speech embeddings for the same emotion class. We also propose a consistency loss to regulate the agreement between expert predictions — each trained with focal loss for class imbalance. The following are the contributions from the work:

- We propose MiSTER-E—a modular framework for ERC that separates modality-specific context modeling from cross-modal fusion using a Mixture-of-Experts architecture.

- We fine-tune LLMs for speech and text, model conversational dynamics via temporal inception networks and BiGRUs, and integrate predictions using a gating mechanism.
- We introduce a speech-text contrastive loss to enhance cross-modal alignment for utterances belonging to the same emotion class. Further, to promote agreement between the experts we introduce a KL-divergence based consistency regularization loss.
- Our proposed method is shown to achieve state-of-the-art performance on two standard ERC benchmarks—IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019a).

2 Related Work

Text embedding extraction: Early approaches to ERC relied on static word embeddings such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), to encode utterances (Poria et al., 2015; Zadeh et al., 2017; Mai et al., 2019). With the advent of transformer-based language models like BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), ERC systems began adopting contextual encoders (Hazarika et al., 2020; Chudasama et al., 2022; Hu et al., 2023), resulting in improved text features. Recent text-only methods (Lei et al., 2023; Fu, 2024) pose ERC as a generative task, where they fine-tune LLMs in an autoregressive manner. However, the efficient use of LLMs when text is used alongside speech is unexplored. Towards this, we adapt LLMs as text encoders for the task of text emotion recognition, thereby harnessing the power of these models and also enabling fusion with other modalities.
Speech embedding extraction: Speech features in ERC have traditionally relied on hand-crafted descriptors like OpenSMILE (Eyben et al., 2010) and COVAREP (Degottex et al., 2014), which, while effective, often fail to generalize across datasets with diverse acoustic conditions (Majumder et al., 2019; Poria et al., 2015). Recent efforts have moved toward learnable frontends such as LEAF (Zeghidour et al., 2021) and self-supervised models like HuBERT (Hsu et al., 2021) and wav2vec (Baevski et al., 2020), with demonstrated success (Dutta and Ganapathy, 2022; Lian et al., 2022). However, the utilization of multi-modal LLMs for ERC is relatively unexplored. Thus, we fine tune large speech language models (SLLMs) directly for emotional inference—an approach that has not been explored for multimodal ERC before.
MoE in ERC: While Mixture-of-Experts (MoE) architectures have shown promise in scaling both language (Shazeer et al., 2017; Lepikhin et al., 2021) and vision models (Riquelme et al., 2021), their role in ERC has been limited. One recent example, MMGAT-EMO (Zhang et al., 2025), combines MoE with graph attention for emotion modeling. In contrast, we use MoE to structure our model around three specialized experts—speech, text, and fused modalities—explicitly targeting the separation of context modeling and cross-modal fusion.
Loss functions for ERC: Supervised contrastive learning (Khosla et al., 2020) was introduced for ERC by Li et al.(Li et al., 2022) and extended in later works (Song et al., 2022; Yu et al., 2024), using emotion class prototypes. Some approaches extend this to align modalities—e.g., aligning audio and visual cues to textual anchors (Hu et al., 2022b). In contrast, we adopt a multimodal supervised contrastive loss, where positives are intra- and inter-modality representations of utterances belonging to the same emotion classes, encouraging better alignment across modalities for each of the emotion categories. We further introduce a consistency loss to encourage agreement among experts, reinforcing modular cooperation.

3 Proposed Method

A block diagram of our proposed method is shown in Fig. 1.

3.1 Problem Description

Let us consider an ERC dataset $\mathcal{D}$ consisting of P conversations, $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \ldots, \mathbf{c}_P\}$, where each conversation $\mathbf{c}_i$ consists of a set of utterances, $\mathbf{U}_i = \{\mathbf{u}_{i1}, \mathbf{u}_{i2}, \ldots, \mathbf{u}_{iN}\}$. In this work, only the speech and text modalities are considered, which is notated as $\mathbf{u}_{ik} = \{\mathbf{s}_{ik}, \mathbf{t}_{ik}\}$, $k = 1..N$, where $\mathbf{s}_{ik}$ and $\mathbf{t}_{ik}$ are speech and text data respectively. Each utterance $\mathbf{u}_{ik}$ is associated with a corresponding emotion label $y_{ik} \in \mathcal{Y}$, with $\mathcal{Y}$ denoting the label set of emotion categories in $\mathcal{D}$. The task of ERC is to map a sequence of utterances $\{\mathbf{u}_{i1}, \ldots, \mathbf{u}_{iN}\}$ to their corresponding labels $\mathbf{y}_i = \{y_{i1}, y_{i2}, \ldots, y_{iN}\}$.

3.2 Unimodal Feature Extraction

Text embeddings: While large language models (LLMs) excel in text generation, their use in emo-

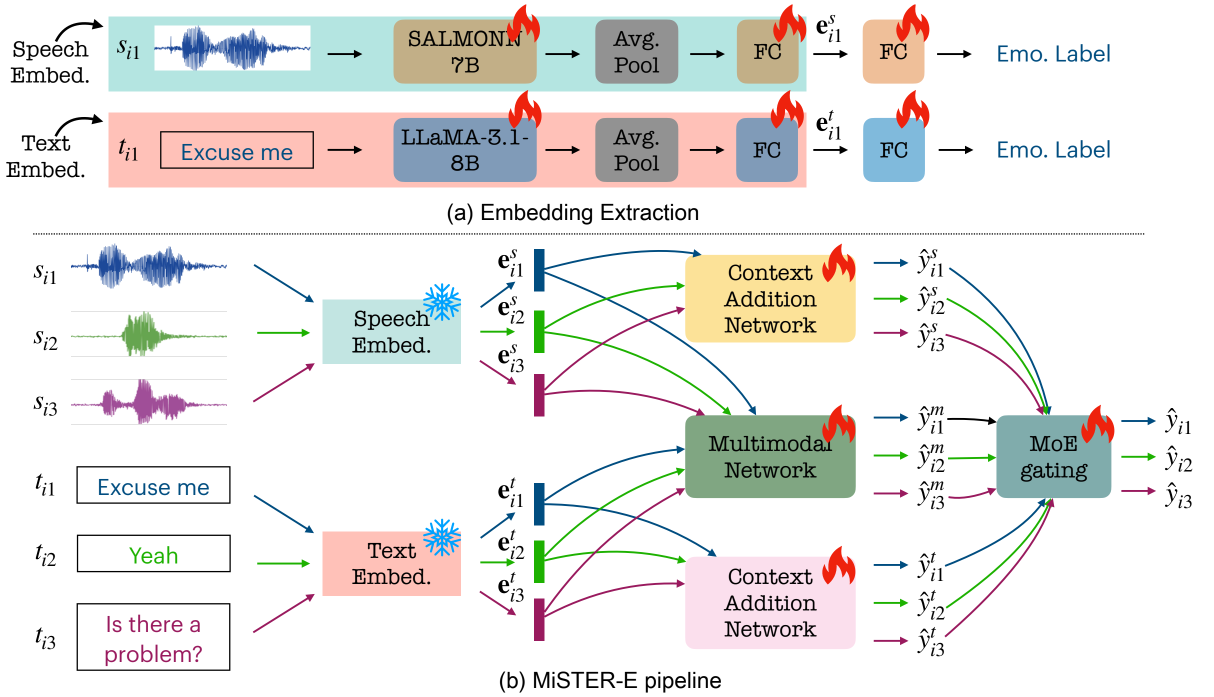


Figure 1: (a) Training of the unimodal feature extraction module (Sec. 3.2) (b) The entire pipeline of MiSTER-E. The speech and text embedding modules are frozen during training of the rest of the pipeline. Two context addition networks (Sec. 3.3.1) are trained for the two modalities along with a multimodal network (Sec. 3.3.2). Finally, a mixture of experts gating network (Sec. 3.4) is trained to predict the emotion category for each utterance.

tion recognition has been limited. Prior work typically uses them in autoregressive mode or as frozen feature extractors (BehnamGhader et al., 2024). In this work, we fine-tune an LLM—specifically, `LLaMA-3.1-8B`—to act as an text encoder rather than a generator. Each utterance transcript $\mathbf{t}_{ik}$ is tokenized and processed through the LLM. Token-level hidden states are then pooled and passed through a two-layer feedforward classifier trained with task-specific supervision. To preserve the pre-training knowledge while enabling efficient adaptation, we apply LoRA (Hu et al., 2022a) to fine-tune the weights. We denote the resulting text embedding as:

$$\mathbf{e}^t_{ik} = \texttt{Text-Embed}(\mathbf{t}_{ik}) \quad (1)$$

where $\mathbf{e}^t_{ik}$ is extracted from the first fully connected layer of the classifier.

Speech embeddings: For speech, we adopt a similar approach using `SALMONN-7B` (Tang et al., 2024), a speech large language model (SLLM) comprising a speech encoder, Q-former (Li et al., 2023b), and an LLM backbone. For ERC, we fine-tune this model by updating the Q-former, the LLM, and a classification head via LoRA. This allows the system to learn emotionally salient acoustic patterns while retaining the semantic features of the LLM backbone. Given a speech signal $\mathbf{s}_{ik}$, we extract its representation as:

$$\mathbf{e}^s_{ik} = \texttt{Speech-Embed}(\mathbf{s}_{ik}) \quad (2)$$

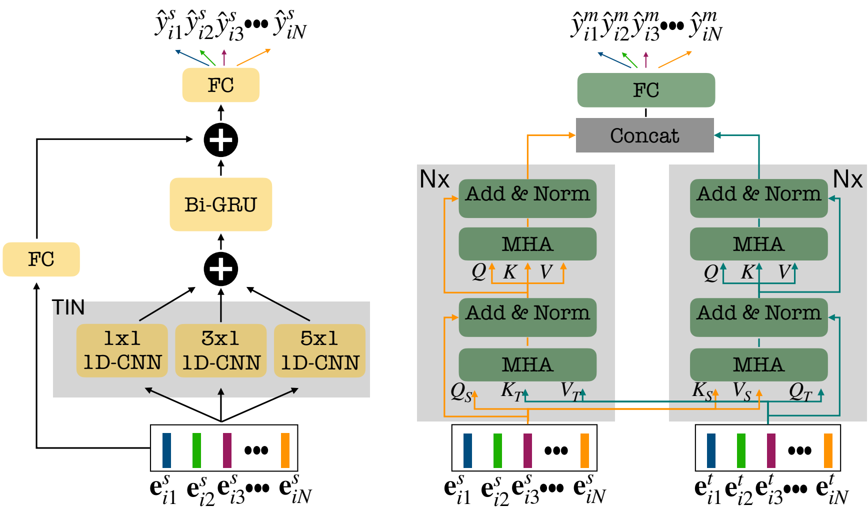
where the embedding is taken from the first fully connected layer of the speech classification head. Comparisons with alternative LLM and SLLM variants are provided in Appendix A.8.

3.3 Conversational Modeling

For a conversation $\mathbf{c}_i$, the text and speech embedding sequences (Sec. 3.2) are denoted by $\mathbf{E^t_i} = \{\mathbf{e}^t_{i1}, \ldots, \mathbf{e}^t_{iN}\}$ and $\mathbf{E^s_i} = \{\mathbf{e}^s_{i1}, \ldots, \mathbf{e}^s_{iN}\}$ respectively. As shown in Figure 1, the fine-tuned speech and text embedding extractors are frozen.

3.3.1 Context Addition Network

To make utterance representations context-aware, we introduce a Context Addition Network (`CAN`) that enhances both text and speech embeddings (Sec. 3.2) with conversational context. At its core is a Temporal Inception Network (`TIN`), which applies 1D convolutions with kernel sizes of 1, 3, and 5 to capture short-range dependencies—simulating varying receptive fields across local utterance neighborhoods. Inspired by the Inception architecture (Szegedy et al., 2015) originally developed



(a) Context addition network (b) Multimodal network

Figure 2: (a) The context addition network (for the speech modality) and (b) the multimodal network used in MiSTER-E. The inputs to both the blocks are derived from the uni-modal feature extractor modules. `TIN` stands for Temporal Inception Network, `MHA` stands for multi-head attention.

for image classification, this is, to the best of our knowledge, the first application of an inception-style network for contextual modeling in ERC.
However, emotional signals in dialogue often evolve over longer durations. To capture such global dependencies, we append a Bi-GRU layer, allowing the model to integrate information across the entire conversational span. A residual connection links the original embedding with its context-enhanced version, enabling additive refinement while preserving the semantic grounding of the base LLM features. Finally, a FC layer is used to map each utterance $\mathbf{u}_{ik}$ to its corresponding emotion category y_{ik}. We train two similar networks for the speech and text modalities, respectively. These operations are denoted as:

$$\hat{\mathbf{y}}_{\mathrm{i}}^{\mathrm{s}} = \{\hat{y}_{i1}^{s}, \ldots, \hat{y}_{iN}^{s}\} = \mathsf{CAN}(\mathbf{E}_{\mathrm{i}}^{\mathrm{s}}) \quad (3)$$
$$\hat{\mathbf{y}}_{\mathrm{i}}^{\mathrm{t}} = \{\hat{y}_{i1}^{t}, \ldots, \hat{y}_{iN}^{t}\} = \mathsf{CAN}(\mathbf{E}_{\mathrm{i}}^{\mathrm{t}}) \quad (4)$$

A schematic of the speech-side `CAN` is shown in Fig. 2(a). Detailed ablation results and performance of an alternative attention-based architecture are discussed in Appendix A.9.

3.3.2 Multimodal Network

To integrate information across modalities, we design a fusion module based on cross-attention between speech and text. This mechanism allows the model to align semantic cues from text with affective signals from audio. The speech and text embeddings, $\mathbf{E}_{\mathrm{i}}^{\mathrm{s}}$ and $\mathbf{E}_{\mathrm{i}}^{\mathrm{t}}$, are first projected into query, key, and value spaces. Cross-attention is then applied bidirectionally:

$$\mathbf{M}_{\mathrm{i}}^{\mathrm{t}\to\mathrm{s}} = \mathsf{LN}(\mathsf{FC}(\mathbf{E}_{\mathrm{i}}^{\mathrm{s}}) + \mathsf{MHA}(\mathbf{Q}_{\mathrm{i}}^{\mathrm{s}}, \mathbf{K}_{\mathrm{i}}^{\mathrm{t}}, \mathbf{V}_{\mathrm{i}}^{\mathrm{t}})) \quad (5)$$
$$\mathbf{M}_{\mathrm{i}}^{\mathrm{s}\to\mathrm{t}} = \mathsf{LN}(\mathsf{FC}(\mathbf{E}_{\mathrm{i}}^{\mathrm{t}}) + \mathsf{MHA}(\mathbf{Q}_{\mathrm{i}}^{\mathrm{t}}, \mathbf{K}_{\mathrm{i}}^{\mathrm{s}}, \mathbf{V}_{\mathrm{i}}^{\mathrm{s}})) \quad (6)$$

Here, `MHA` and `LN` refer to multi-head attention and layer normalization, respectively.
While this enables inter-modal alignment, it overlooks cues from the conversational context that are essential for ERC. To address this, we apply modality-specific self-attention layers over the aligned representations, capturing temporal context across the conversation (See Appendix A.10):

$$\mathbf{M}_{\mathrm{i}}^{\mathrm{s}} = \texttt{Self-attn.}(\mathbf{M}_{\mathrm{i}}^{\mathrm{t}\to\mathrm{s}}) \quad (7)$$
$$\mathbf{M}_{\mathrm{i}}^{\mathrm{t}} = \texttt{Self-attn.}(\mathbf{M}_{\mathrm{i}}^{\mathrm{s}\to\mathrm{t}}) \quad (8)$$

The outputs $\mathbf{M}_{\mathrm{i}}^{\mathrm{s}}$ and $\mathbf{M}_{\mathrm{i}}^{\mathrm{t}}$ are concatenated and passed through a fully connected (FC) layer:

$$\hat{\mathbf{y}}_{\mathrm{i}}^{\mathrm{m}} = \{\hat{y}_{i1}^{m}, \hat{y}_{i2}^{m}, \ldots, \hat{y}_{iN}^{m}\} = \mathsf{FC}([\mathbf{M}_{\mathrm{i}}^{\mathrm{s}}; \mathbf{M}_{\mathrm{i}}^{\mathrm{t}}]) \quad (9)$$

An overview is presented in Fig. 2(b).

3.4 Mixture-of-Experts Gating

We obtain outputs from the three experts: the speech expert ($\hat{\mathbf{y}}_{\mathrm{i}}^{\mathrm{s}}$), the text expert ($\hat{\mathbf{y}}_{\mathrm{i}}^{\mathrm{t}}$), and the multimodal expert ($\hat{\mathbf{y}}_{\mathrm{i}}^{\mathrm{m}}$).
To combine the outputs, we dynamically fuse the experts' decisions using a gating mechanism that learns to weigh the decisions based on its confidence for the given utterance. To compute these

weights, we concatenate the expert predictions and feed them into a fully connected (FC) layer:

$$\mathbf{g}_i = \mathsf{FC}([\hat{\mathbf{y}}_{\mathbf{i}}^{\mathbf{s}}; \hat{\mathbf{y}}_{\mathbf{i}}^{\mathbf{t}}; \hat{\mathbf{y}}_{\mathbf{i}}^{\mathbf{m}}]) \tag{10}$$

The outputs from the gating network, $\mathbf{g}_i \in \mathbb{R}^{N\times 3}$, are transformed via a softmax operation to produce adaptive mixture weights $\beta_i = [\beta_i^s, \beta_i^t, \beta_i^m]$, which indicate the relative importance of each expert. The final prediction is computed as a weighted sum of the expert predictions:

$$\hat{\mathbf{y}}_{\mathbf{i}} = \beta_i^s \cdot \hat{\mathbf{y}}_{\mathbf{i}}^{\mathbf{s}} + \beta_i^t \cdot \hat{\mathbf{y}}_{\mathbf{i}}^{\mathbf{t}} + \beta_i^m \cdot \hat{\mathbf{y}}_{\mathbf{i}}^{\mathbf{m}} \tag{11}$$

This gating network is trained end-to-end, empowering the model to flexibly integrate modalities by emphasizing the most reliable expert.

3.5 Model Training

3.5.1 Loss Function

ERC is typically characterized by severe class imbalance, where rare emotion classes are often misclassified (Poria et al., 2019b). To address this, we employ the focal loss (Lin et al., 2017), which modulates the contribution of each training example based on its complexity, reducing the relative loss for well-classified examples while focusing on hard (possibly minority class) instances. In MiSTER-E, this loss is applied during the training of the uni-modal embedding extractors for text and speech as well as their respective context addition networks (CANs). The loss for the speech and text CAN networks is given by:

$$\mathcal{L}_{\text{CAN}}^i = \sum_{k=1}^{N} \mathsf{FL}(\hat{y}_{ik}^s, y_{ik}) + \sum_{k=1}^{N} \mathsf{FL}(\hat{y}_{ik}^t, y_{ik}) \tag{12}$$

Here, $\mathsf{FL}(\cdot)$ represents the focal loss function. Description of the focal loss and related ablations are provided in Appendix. A.1.

3.5.2 Multimodal Contrastive Loss

In many cases, speech and text provide complementary signals about a speaker's emotion. To exploit this, we incorporate a supervised contrastive loss that structures the joint representation space based on emotion labels, drawing together utterances with shared emotional intent across modalities. Consider the multimodal speech and text representations denoted by $\mathbf{M}_{\mathbf{i}}^{\mathbf{s}} = \{\mathbf{m}_{\mathbf{i1}}^{\mathbf{s}}, \mathbf{m}_{\mathbf{i2}}^{\mathbf{s}}, \mathbf{m}_{\mathbf{i3}}^{\mathbf{s}}, \ldots, \mathbf{m}_{\mathbf{iN}}^{\mathbf{s}}\}$ and $\mathbf{M}_{\mathbf{i}}^{\mathbf{t}} = \{\mathbf{m}_{\mathbf{i1}}^{\mathbf{t}}, \mathbf{m}_{\mathbf{i2}}^{\mathbf{t}}, \mathbf{m}_{\mathbf{i3}}^{\mathbf{t}}, \ldots, \mathbf{m}_{\mathbf{iN}}^{\mathbf{t}}\}$ respectively. These embeddings are batched to get,

$$\mathbf{Z}_{\mathbf{i}} = \{\tilde{\mathbf{m}}_{\mathbf{i1}}^{\mathbf{s}}, \ldots, \tilde{\mathbf{m}}_{\mathbf{iN}}^{\mathbf{s}}, \tilde{\mathbf{m}}_{\mathbf{i1}}^{\mathbf{t}}, \ldots, \tilde{\mathbf{m}}_{\mathbf{iN}}^{\mathbf{t}}\} \tag{13}$$

where $\tilde{\mathbf{m}}_{\mathbf{ik}} = \frac{\mathbf{m}_{\mathbf{iN}}}{||\mathbf{m}_{\mathbf{iN}}||}$ denotes the normalized embeddings. Let $\mathbf{z}_a \in \mathbf{Z}_{\mathbf{i}}$ for $a \in \{1, 2, \ldots, 2N\}$, and let y_a be the emotion label associated with $\mathbf{z}_a$. The contrastive loss is given by:

$$\mathcal{L}_{\text{con}}^{\text{i}} = \sum_{a=1}^{2N} \frac{-1}{|P(a)|} \sum_{p\in P(a)} \log \frac{\exp\left(\mathbf{z}_a^\top \mathbf{z}_p/\tau\right)}{\sum_{\substack{q=1\\q\neq a}}^{2N} \exp\left(\mathbf{z}_a^\top \mathbf{z}_q/\tau\right)} \tag{14}$$

where τ is the temperature and $P(a) = \{p \in \{1, 2, \ldots, 2N\}\setminus\{a\} | y_p = y_a\}$ is the set of positives for anchor $\mathbf{z}_a$. This objective pulls together utterances with the same emotion class across both speech and text, while pushing apart other samples, thereby guiding the model to learn emotionally coherent, modality-invariant representations.

Total multimodal loss: The final loss used to train the multimodal model combines the classification and contrastive objectives:

$$\mathcal{L}_{\text{multi}}^i = \sum_{k=1}^{N} \mathsf{FL}(\hat{y}_{ik}^m, y_{ik}) + \lambda \mathcal{L}_{\text{con}}^i \tag{15}$$

where $\hat{y}_{ik}^m$ denotes the multimodal prediction, and λ controls the contribution of the contrastive loss.

3.5.3 MoE Gating Loss

To train the mixture-of-experts (MoE) gating network, we combine two objectives: (i) focal loss for emotion classification, and (ii) a regularization term to promote consistency among the expert predictions. Specifically, we enforce similarity in the predicted distributions of the three experts—speech-only, text-only, and multimodal—using the Kullback-Leibler (KL) divergence. The total loss for the MoE layer is:

$$\mathcal{L}_{\text{moe}}^i = \sum_{k=1}^{N} \mathsf{FL}(\hat{y}_{ik}, y_{ik}) + \alpha \cdot \mathcal{L}_{\text{KL}}^i \tag{16}$$

where $\hat{y}_{ik}$ is defined in Eq. 11, α controls the strength of the consistency regularization, and the KL term is given by:

$$\mathcal{L}_{\text{KL}}^i = \sum_{k=1}^{N} \Big[\mathsf{KL}(\hat{y}_{ik}^m | \hat{y}_{ik}^s) + \mathsf{KL}(\hat{y}_{ik}^m | \hat{y}_{ik}^t) + \mathsf{KL}(\hat{y}_{ik}^s | \hat{y}_{ik}^t)\Big] \tag{17}$$

This encourages the expert branches to produce aligned output distributions, enabling the gating mechanism to combine them effectively.

Method	Modalities Used	IEMOCAP	MELD
bc-LSTM (Poria et al., 2017)	T,S,V	54.9%	55.9%
UniMSE (Hu et al., 2022b)	T,S,V	70.7%	65.5%
SCMM (Yang et al., 2023)	T,S,V	67.5%	59.4%
GraphSmile (Li et al., 2024b)	T,S,V	72.8%	66.7%
SMIN (Lian et al., 2022)##	T,S	70.5%	63.7%
MultiEmo (Shi and Huang, 2023)#	T,S	$66.9\%^{\pm 2.0}$	$65.3\%^{\pm 0.5}$
HCAM (Dutta and Ganapathy, 2023)	T,S	70.5%	65.8%
DF-ERC (Li et al., 2023a)	T,S	69.5%	64.5%
Mamba-like-model (Shou et al., 2024)	T,S	70.2%	65.6%
CFN-ESA (Li et al., 2024a)	T,S	68.7%	67.2%
MMGAT-EMO (Zhang et al., 2025)#	T,S	$65.5\%^{\pm 0.6}$	$66.1\%^{\pm 0.4}$
MiSTER-E	T,S	$\mathbf{70.9}\%^{\pm 0.2}$	$\mathbf{69.5}\%^{\pm 0.3}$

Table 1: Comparison of different methods on IEMOCAP and MELD datasets on weighted-F1 scores. We mention the modalities used by the methods (T:Text, S:Speech, V:Video). Further we compare with only those methods which do not use any speaker information. ## we report the numbers when no external emotional data is used for training SMIN. # we ran the public implementation provided by the authors on the datasets for our own settings. The superscript results are the mean and standard deviation over 3 random initializations, whenever performed.

3.5.4 Total Loss

The context addition networks, the multimodal network, and the MoE gating layer are trained together. The total loss is:

$$\mathcal{L}_{\text{tot}} = \sum_{i=1}^{P} \left[\mathcal{L}^i_{\text{CAN}} + \mathcal{L}^i_{\text{moe}} + \mathcal{L}^i_{\text{multi}} \right] \tag{18}$$

4 Experiments

4.1 Datasets

We evaluate the proposed method on two benchmark ERC datasets - IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019a). More details are available in Appendix A.2.

IEMOCAP consists of conversational data split into 5 sessions, 151 dialogues and 7433 utterances. Following prior work (Lian et al., 2022), we consider session 5 for testing, while session 1 is used for validation. The remaining 3 sessions are used for training, which is identical to the setup followed in prior works. Each utterance is classified as one of six emotions: “angry”, “happy”, “sad”, “frustrated”, “excited” and “neutral”.

MELD is a multi-party conversational dataset consisting of 1433 dialogues and 13708 utterances from the TV show *Friends*. This dataset has predefined train, validation, and test splits which are used in this work. Each utterance is categorized as one of seven emotion classes: “angry”, “joy”, “sadness”, “fear”, “disgust”, “surprise” and “neutral”.

4.2 Implementation details

The two unimodal feature extractors (`SALMONN-7B` and `LLaMA-3.1-8B`) are trained using LoRA with a rank of 8 and the scaling parameter of 32 with a dropout of 0.1. Both models are trained with a batch size of 8 and a learning rate of $1e-5$, with the focal loss. The hidden dimension in the FC (Sec. 3.2) is set to 2048. For the rest of the MiSTER-E pipeline, we use a batch size of 8 for IEMOCAP and 32 for MELD. The learning rate is set to $1e-5$ for both datasets. For MELD, λ (Eq. 15) is set to 1 , while for IEMOCAP it is set to 2. The consistency regularization term, α (Eq. 16), is set to 0.1 for IEMOCAP, while it is kept at $1e-3$ for MELD (See Appendix A.12). We report the weighted F1-score on the test data as the performance metric with 3 random initializations. More hyperparameter and implementation details are given in Appendix A.3. Code and trained models will be made public upon acceptance.

4.3 Comparison with prior work

We compare the proposal with several baseline approaches, described in Appendix A.4. While many existing systems leverage all three modalities, we focus on comparisons with those using only speech and text. We note that for IEMOCAP, due to the lack of a standardized validation set, model selection for many of the prior works is often performed on the test set (see Appendix A.6 for discussion on this aspect). Some of the prior works also exploit

Method	# Params	Text Feats.	Speech Feats.	IEMOCAP	MELD
MultiEmo (Shi and Huang, 2023)	≈ 450M	RoBERTa	OpenSMILE	66.9%	65.3%
+ LLM/SLLM	≈ 14B	LLaMA	SALMONN	66.5%	68.2%
HCAM (Dutta and Ganapathy, 2023)	≈ 750M	RoBERTa	wav2vec	<u>70.5%</u>	65.8%
+ LLM/SLLM	≈ 14B	LLaMA	SALMONN	70.3%	<u>68.4%</u>
MMGAT-EMO (Zhang et al., 2025)	≈ 430M	EmoBERTa	OpenSMILE	65.5%	66.1%
+ LLM/SLLM	≈ 14B	LLaMA	SALMONN	66.9%	66.1%
MiSTER-E	≈ 14B	LLaMA	SALMONN	**70.9%**	**69.5%**

Table 2: Comparison of some of the baseline methods when re-designed with LLM features.

speaker information, and we analyze this effect in Appendix A.5.

Table 1 reports the performance of MiSTER-E and other prior approaches. It is seen that our proposed method achieves state-of-the-art performance among other prior works, with a weighted F1 score of **70.9%** on IEMOCAP and **69.5%** on MELD. Detailed class-wise results are given in Appendix A.13 and a case study is reported in A.14.

Recent LLM-based text-only ERC systems (e.g., InstructERC (Lei et al., 2023), CKERC (Fu, 2024), BiosERC (Xue et al., 2024)) incorporate speaker roles or external knowledge, making direct comparison to MiSTER-E challenging. However, we include a controlled comparison on a text-only dataset in Appendix A.7.

4.4 Why is modularity crucial?

We replace the modular design of our proposal with a monolithic architecture that feeds the contextual representations into the fusion module without separating the experts. This results in a significant performance drop: from 70.9% to 67.8% on IEMOCAP, and from 69.5% to 67.9% on MELD. These results confirm that disentangling context modeling from multimodal fusion—as done in MiSTER-E—is not only conceptually sound but also empirically crucial for effective emotion recognition.

4.5 Are LLM embeddings the panacea?

One might ask whether the gains of our approach stem solely from the use of LLM/SLLM features. Towards this, we replace the original features in several prior models with the same LLM/SLLM embeddings used in our method and retrain them (Table 2). While MELD sees modest gains, IEMOCAP is impacted marginally (or even negatively)—suggesting that LLM representations alone are not sufficient. Notably, MiSTER-E still outperforms all baselines, underscoring that its effectiveness lies not just in LLM features, but in its architectural design that separates context modeling from multimodal fusion.

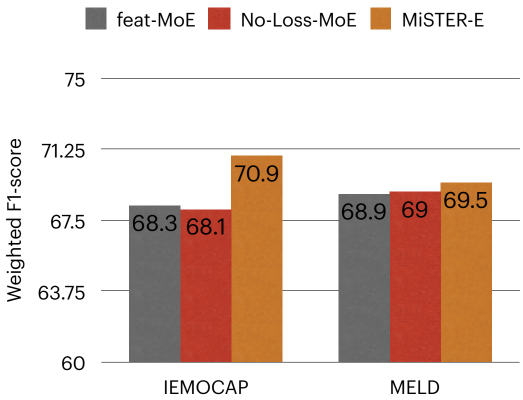


Figure 3: Performance of MiSTER-E with changes in the MoE gating strategy for IEMOCAP and MELD.

4.6 Is decision-level MoE gating crucial?

To evaluate our gating strategy, we compare two architectural variants. First, instead of fusing expert predictions at the decision level, we perform Mixture-of-Experts fusion at the feature level (`feat-MoE`) before classification. Second, we remove expert-specific supervision by training only on the gated output with a focal loss (`No-Loss-MoE`), omitting individual losses for the audio, text, and multimodal experts. As shown in Fig. 3, both variants degrade performance: `feat-MoE` results in a 2.6% drop on IEMOCAP and 0.6% on MELD, while `No-Loss-MoE` leads to similar degradation. These results confirm that decision-level fusion preserves modality-specific discriminative cues, and that expert-level training is essential for enabling each expert to specialize.

5 Discussion

5.1 Importance of the Contrastive Loss

We analyze the impact of the contrastive loss (Eq. 14) on model performance in Fig. 4.

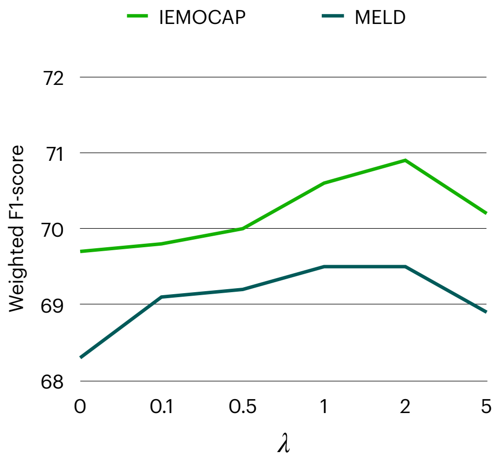


Figure 4: Performance of MiSTER-E with varying values of λ (Eq. 18) for the two datasets. The validation set performance (not shown above) mirrors the trends of the test set. For IEMOCAP, validation set performance is 63% for $\lambda = 0$ and 63.6% for $\lambda = 2$. For MELD, the highest validation set performance is achieved for $\lambda = 1$ (65.6%) as compared to 65.4% for $\lambda = 0$.

Key Takeaways: 1) On MELD, performance improves from 68.2% (with λ=0) to 69.5% at λ=1, while IEMOCAP experiments suggest an improvement from 70.2% to 70.9% with $\lambda = 2$. This suggests that the contrastive loss enhances the ability of the fusion module to align emotion-relevant cues across modalities beyond what focal loss can achieve alone. 2) Increasing the value of λ is seen to degrade the performance. This indicates that, although the proposed contrastive loss aids in classification, it leads to a drop in performance without the focal loss.

5.2 Analysis of Expert Weights

We analyze the expert weights assigned by the MoE gating network on IEMOCAP and MELD (Fig. 5). **Key Takeaways**: 1) The gating mechanism exhibits dataset-specific preferences—favoring the multimodal expert for IEMOCAP and the text expert for MELD—demonstrating its adaptability to dataset characteristics (see Appendix A.11 for individual expert performance). 2) MELD displays higher variance in expert weights, likely reflecting its greater variability compared to the more controlled IEMOCAP dataset. 3) A case study from MELD (Fig. 6) illustrates dynamic expert selection: for a four-utterance conversation with distinct emotions per turn. The gating prioritizes speech and multimodal experts for the first utterance, shifting

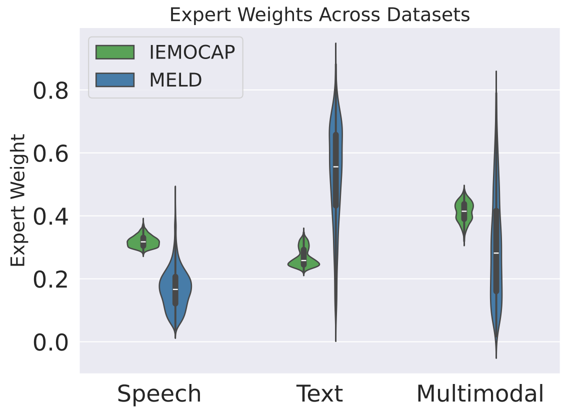


Figure 5: Distribution of weights for the experts for the different datasets.

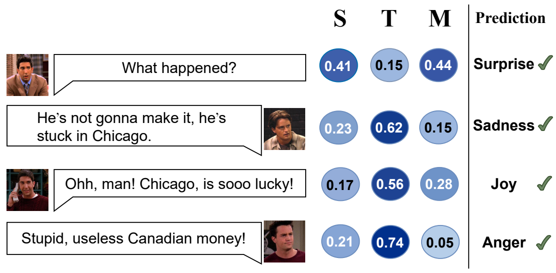


Figure 6: Distribution of weights for an example conversation from MELD. S: Speech expert, T: Text expert, M: Multimodal expert. The predictions by MiSTER-E are also shown. Refer Fig. 12 for another example.

to the text expert for subsequent utterances. This highlights the model's ability to adapt the emphasis on the experts at a fine-grained, per-utterance level.

6 Conclusion

We proposed MiSTER-E, a modular framework for ERC that explicitly separates contextual modeling from multimodal fusion. Leveraging LLM-based representations for both speech and text, we model context in the conversations using a temporal inception block followed by a Bi-GRU, and perform modality fusion via an attention-based network. A Mixture-of-Experts (MoE) gate adaptively integrates decisions from context-aware and multimodal experts. MiSTER-E achieves new state-of-the-art performance on IEMOCAP and MELD, without using speaker identity, demonstrating the efficacy of modular context-fusion and decision-level gating. The different design choices are further justified by means of extensive ablations.

7 Limitations

While MiSTER-E achieves strong performance, some limitations remain. First, the LLM/SLLM encoders introduce some computational overhead. Second, although we avoid explicit speaker identity modeling, some prior works suggest that speaker-aware models can capture interpersonal dynamics more effectively. Using the speaker information in an unsupervised way is an avenue unexplored in this work. Finally, we focus exclusively on the speech and text modalities, omitting the visual channel available in datasets like IEMOCAP and MELD. Integrating visual cues could further enhance model performance.

References

Wei Ai, Fuchen Zhang, Yuntao Shou, Tao Meng, Haowen Chen, and Keqin Li. 2025. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11418–11426.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4652–4661.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Soumya Dutta and Sriram Ganapathy. 2022. Multimodal transformer with learnable frontend and self attention for emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6917–6921. IEEE.

Soumya Dutta and Sriram Ganapathy. 2023. Hcam–hierarchical cross attention model for multimodal emotion recognition. *arXiv preprint arXiv:2304.06910*.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Yumeng Fu. 2024. Ckerc: Joint large language models with commonsense knowledge for emotion recognition in conversation. *arXiv preprint arXiv:2403.07260*.

Bharat Gaind, Varun Syal, and Sneha Padgalwar. 2019. Emotion detection and analysis on social media. *arXiv preprint arXiv:1901.08458*.

Surjya Ghosh, Sumit Sahu, Niloy Ganguly, Bivas Mitra, and Pradipta De. 2019. Emokey: An emotion-aware smartphone keyboard for mental health monitoring. In *2019 11th international conference on communication systems & networks (COMSNETS)*, pages 496–499. IEEE.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10835–10852.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022a. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022b. 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.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *arXiv preprint arXiv:2309.11911*.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*.

Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. 2023a. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5923–5934.

Jiang Li, Xiaoping Wang, Yingjian Liu, and Zhigang Zeng. 2024a. Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. *IEEE Transactions on Affective Computing*.

Jiang Li, Xiaoping Wang, and Zhigang Zeng. 2024b. Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition. *arXiv preprint arXiv:2407.21536*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make bart a good dialogue emotion recognizer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11002–11010.

Zheng Lian, Bin Liu, and Jianhua Tao. 2022. Smin: Semi-supervised multi-modal interaction network for conversational emotion recognition. *IEEE Transactions on Affective Computing*, 14(3):2415–2429.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 481–492.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Maja Pantic, Nicu Sebe, Jeffrey F Cohn, and Thomas Huang. 2005. Affective multimodal human-computer interaction. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 669–676.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2539–2544.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*.

Siyuan Shen, Feng Liu, Hanyang Wang, and Aimin Zhou. 2025. Towards speaker-unknown emotion recognition in conversation via progressive contrastive deep supervision. *IEEE Transactions on Affective Computing*.

Tao Shi and Shao-Lun Huang. 2023. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766.

Yuntao Shou, Tao Meng, Fuchen Zhang, Nan Yin, and Keqin Li. 2024. Revisiting multi-modal emotion learning with broad state space models and probability-guidance fusion. *CoRR*.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, MA Zejun, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. 2024. Bioserc: Integrating biography speakers supported by llms for erc tasks. In *International Conference on Artificial Neural Networks*, pages 277–292. Springer.

Haozhe Yang, Xianqiang Gao, Jianlong Wu, Tian Gan, Ning Ding, Feijun Jiang, and Liqiang Nie. 2023. Self-adaptive context and modal-interaction modeling for multimodal emotion recognition. In *Findings of the association for computational linguistics: ACL 2023*, pages 6267–6281.

Fangxu Yu, Junjie Guo, Zhen Wu, and Xinyu Dai. 2024. Emotion-anchored contrastive learning framework for emotion recognition in conversation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4521–4534.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *AAAI Workshops*, volume 18, pages 44–52.

Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. 2021. Leaf: A learnable frontend for audio classification. In *International Conference on Learning Representations*.

Chengwen Zhang, Yaohui Liu, and Bo Cheng. 2025. A moe multimodal graph attention network framework for multimodal emotion recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A Appendix

A.1 Focal Loss

Focal loss (Lin et al., 2017) is generally used to handle class imbalance. Since both ERC datasets considered in this paper are imbalanced, we use focal loss in place of the cross-entropy loss.

Denoting the predicted probability for the sample with logits $\hat{y}^{\cdot}_{ik}$ as $\hat{p}^{\cdot}_{ik}$, the focal loss is given as,

$$\mathsf{FL}(\hat{y}^{\cdot}_{ik}, y_{ik}) = -\sum_{c=1}^{\mathcal{Y}} y^{c}_{ik}(1-\hat{p}^{\cdot c}_{ik})^{\gamma}\log(\hat{p}^{\cdot c}_{ik}) \quad (19)$$

where $y^c_{ik} = 1$ for the true class and 0 otherwise. $\hat{p}^{\cdot c}_{ik}$ is the predicted probability for class c for the sample with logits $\hat{y}^{\cdot}_{ik}$ ($\cdot$ is replaced by s, t or m) and γ is the hyperparameter associated with the focal loss.

The performance of our proposed method for different values of γ (Eq. 19) is shown in Fig. 7. We note that for cross-entropy loss ($\gamma = 0$), the performance drops for both datasets, indicating the need for utilizing a loss function more suited for imbalanced datasets. The value of γ is set to 3 for both datasets.

A.2 More details about datasets

For IEMOCAP, we use 92 conversations for training, 28 conversations for validation and the 31 conversations in Session 5 for testing. There are a total of 5810 utterances for training and validation while 1623 utterances are used for testing the model. In the case of MELD, the training and validation data together amount to 1153 conversations (11098 utterances), while the test data comprises 280 conversations (2610 utterances).

A.3 More Implementation details

All our experiments are run using a NVIDIA RTX A6000 GPU card with Pytorch 2.7.0 [1] with CUDA 12.6. The number of trainable parameters in our model amounts to 97M (8M each for the LLM/SLLM LoRA training and 81M for the conversational modeling). Training the LLM takes about 10 minutes per epoch, while the SLLM takes about 20 minutes per epoch. The rest of the model after feature encoders takes approximately 10 minutes for 100 epochs. We mention further hyperparameter choices for both datasets below:

- The BiGRU used in the `CAN` network (Sec. 3.3.1) has a hidden dimension of 512 and we use 3 layers for IEMOCAP and 2 for MELD. We use a dropout of 0.2 between each fully connected layer for regularization.

- For the multimodal fusion network, we use 4 layers with hidden dimension of 120 and 4 attention heads. We use a dropout regularization of 0.5 in the multimodal fusion network. The same network is used for both the datasets.

- The temperature parameter for the contrastive loss (Eq. 14) is kept at 1 for IEMOCAP and 0.05 for MELD.

[1] `https://pytorch.org/blog/pytorch-2-7/`

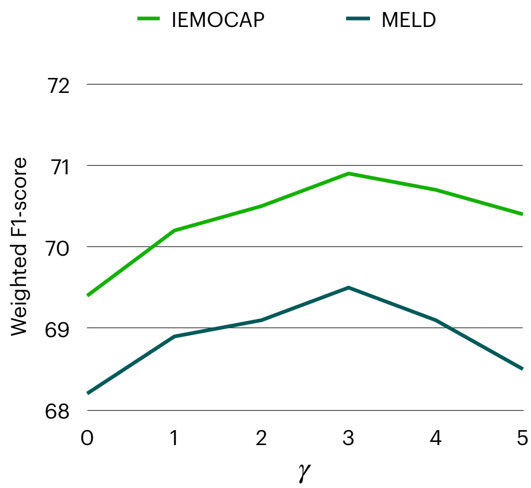


Figure 7: Performance of MiSTER-E with different values of the focal loss hyperparameter.

- While training the context addition networks, the multimodal network, and the MoE gating network, we use gradient clipping with norm 1.0 for both the datasets and train for 100 epochs. While training the LLM and the SLLM feature encoders, we run for 50 epochs for both datasets.

A.4 Baselines

We mention the performance of four prior works using the three modalities-text, speech, and video-for ERC: 1) **bc-LSTM** (Poria et al., 2017): A hierarchical model is proposed that combines modalities by simple concatenation. 2) **UniMSE** (Hu et al., 2022b): The tasks of sentiment and emotion recognition are combined with a contrastive loss between the modality representations with text as the anchor modality. 3) **SCMM** (Yang et al., 2023): The authors design adaptive paths for different modality interactions and hence effective fusion. 4) **GraphSmile** (Li et al., 2024b): Modality interactions are modeled as a graph and sentiment prediction and emotion prediction tasks are combined.
We now mention the baselines using speech and text that we have compared with: 1) **SMIN** (Lian et al., 2022): A semi-supervised learning framework is provided for ERC, where the speech and text representations are reconstructed on a different dataset. 2) **MultiEMO** (Shi and Huang, 2023): A cross-attention fusion method is provided along with a sample weighted focal loss for addressing the class imbalance in the ERC datasets. Using

Method	Modalities used	Speaker Info.	IEMOCAP	MELD
MMGCN	T,S,V	✓	66.2%	58.7%
MMGCN	T,S,V	✗	65.8%	58.4%
SDT	T,S,V	✓	74.1%	66.6%
SDT	T,S,V	✗	72.0%	66.2%
GS-MCC	T,S,V	✓	73.3%	69.0%
GS-MCC	T,S,V	✗	70.6%	64.6%

Table 3: Impact of speaker information in ERC datasets.

their public implementation [2], we train and test this model for both of our datasets and report the numbers. 3) **HCAM** (Dutta and Ganapathy, 2023): A hierarchical model for modeling the different aspects of ERC has been explored - first adding context and then adding the multimodal fusion step. 4) **DF-ERC** (Li et al., 2023a): A technique to disentangle the importance of context and multimodality for ERC is proposed following which the fusion is carried out. 5) **Mamba-like-model** (Shou et al., 2024): State-space models are explored for multimodal ERC for the first time in this paper. 6) **CFN-ESA** (Li et al., 2024a): Cross-attention networks are proposed for fusion of the modalities and the shifting of emotions from one utterance to another is modeled for effective ERC performance. 7) **MMGAT-EMO** (Zhang et al., 2025): A MoE approach is proposed for the multiple modalities in a graph-attention fusion framework. We replicate the results for this method for the datasets using their public implementation [3].

A.5 Speaker Information in ERC

For understanding the role that speaker information plays for ERC, we report the performance of three recent works, **MMGCN** (Hu et al., 2021), **SDT** (Yang et al., 2023) and **GS-MCC** (Ai et al., 2025) with and without speaker information in Table 3. We note that for all these works, the performance drops when the speaker information is not used. The drop is most significant in the case of GS-MCC with a 4.4% drop in the case of MELD. Since MELD does not have a speaker-independent test split, using speaker information in ERC modeling risks an over-estimation of the model performance in the case of MELD.

A.6 Performance on IEMOCAP

IEMOCAP (Busso et al., 2008) is one of the most widely used datasets for evaluating ERC systems. It has conversations split into 5 sessions, and the general evaluation protocol consists of training models on conversations in Session 1 to 4 and testing on Session 5 conversations. However, in the absence of a validation split, we find that many prior works use the test data for validating and saving their models. This often leads to misleading numbers, and hence hurts reproducibility. Thus, for our experiments (wherever we could find a working implementation), we use our train, validation, and test splits, and hence some of our reported numbers for IEMOCAP in Table 1 may vary from those reported in the respective papers.

[2] https://github.com/TaoShi1998/MultiEMO
[3] https://github.com/tdfxlyh/MMGAT_EMO

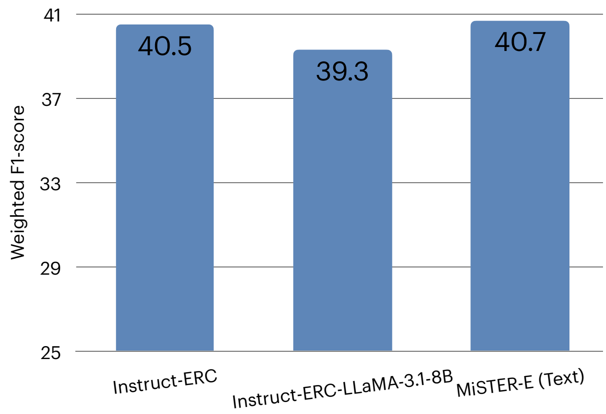


Figure 8: Performance comparison on EmoryNLP dataset.

A.7 Text Only Experiment

MiSTER-E is a speech-text model, suitable for multimodal ERC. However, recently, many methods based on LLMs have been proposed for ERC after framing it as a generative task. While using the text transcripts for fine-tuning the LLMs, they use the speaker information as well. We therefore select a recently proposed method, InstructERC (Lei et al., 2023) and adapt it for our task without using any speaker information for fair comparison with our proposed method. We use the EmoryNLP dataset (Zahiri and Choi, 2018) for this experiment, which consists of 7 emotion classes and has only text transcripts. The text embeddings are extracted as for the other datasets, while the speech embeddings are also replaced by those for the text transcripts. The entire model is trained as before with same hyperparameters as MELD (with $\lambda = 0$ (Eq. 15)). The comparative performance of our proposed method with InstructERC is shown in Fig. 8. We note that our proposed method outperforms InstructERC on this dataset by a slim margin of 0.2%. However, InstructERC adapts a `LLaMA-2-7B-chat`[4] for this task. Since MiSTER-E uses the `LLaMA-3.1-8B` model, we also modify InstructERC by using this model. This system, called Instruct-ERC-LLaMA-3.1-8B (see Fig. 8), is seen to perform worse than both InstructERC and MiSTER-E. This shows the utility of our training methodology over other LLM fine-tuning methods.

A.8 Performance of Other LLMs/SLLMs

We experiment with other SLLMs and LLMs for encoding speech and text, respectively. For all these choices, we train the models with exactly the same hyperparameters as we did for `LLaMA-3.1-8B`[5] and `SALMONN-7B`[6]. For encoding text, we experiment with `Qwen2.5-7B`[7] and `Gemma-7B`[8] while for speech we use `Qwen2-Audio-7B`[9] and `SALMONN-13B`[10]. The results of this experiment are shown in Table 4.

Modality	Model	IEMOCAP	MELD
Text	Gemma-7B	48.2%	58.1%
	Qwen-2.5	53.1%	65.5%
	LLaMA-3.1-8B	**55.3%**	**67.1%**
Speech	SALMONN-7B	**59.7%**	54.3%
	Qwen2-Audio-7B	59.2%	**54.9%**
	SALMONN-13B	58.7%	53.2%

Table 4: Performance of different SLLM/LLMs on the two datasets when the unimodal feature encoders are trained (Sec. 3.2).

We note that for speech, the different speech large language models perform similarly with `SALMONN-7B` performing the best for IEMOCAP and `Qwen2-Audio-7B` outperforming others in the case of MELD. Interestingly, the much bigger `SALMONN-13B` model performs relatively poorly for both datasets. For the text encoder, `LLaMA-3.1-8B` performs the best for both datasets. Based on these results, we choose `SALMONN-7B` and `LLaMA-3.1-8B` for encoding speech and text respectively in MiSTER-E.

[4] https://huggingface.co/meta-llama/Llama-2-7b-chat
[5] https://huggingface.co/meta-llama/Llama-3.1-8B
[6] https://huggingface.co/tsinghua-ee/SALMONN-7B
[7] https://huggingface.co/Qwen/Qwen2.5-7B
[8] https://huggingface.co/google/gemma-7b
[9] https://huggingface.co/Qwen/Qwen2-Audio-7B
[10] https://huggingface.co/tsinghua-ee/SALMONN

A.9 The Context Addition Network

We make the following modifications to the proposed `CAN` network (Fig. 2(a)).

- We remove the temporal inception network (`TIN`) from the network.
- We remove the skip connection between the input and the output of the `CAN` architecture.
- `Bi-GRU+Attn.`: Instead of the TIN network before the Bi-GRU, we append a self-attention block (2 blocks, 120 hidden dimension and 4 heads) after the Bi-GRU for effective contextual modeling.

The weighted F1-scores for the two datasets are shown in Table 5 with these modifications. The different parts of the proposed `CAN` network are seen to be important towards the performance of our proposed method. The local contextual dependencies captured by the temporal inception network are seen to benefit IEMOCAP by 1.1% and MELD by 1.7%. The skip connection is also seen to have a positive impact for both datasets. Interestingly, the combination of Bi-GRU with self-attention is outperformed by `CAN` for both datasets. We notice an overfitting problem in this case - partially explaining this observation.

Method	IEMOCAP	MELD
CAN	70.9%	69.5%
- TIN	69.8%	67.8%
- Skip	69.9%	68.8%
Bi-GRU+Attn.	70.2%	68.5%

Table 5: Performance of MiSTER-E with changes in the `CAN` network.

A.10 Removal of the Self-Attention Layers

We remove the self-attention layers after the cross-attention block in the multimodal network (See Fig. 2(b)). On removal of this block, the performance of our proposed method drops to 70.5% for IEMOCAP and 68.7% for MELD. The drop of 0.4% and 0.8% for IEMOCAP and MELD respectively indicates the utility of the self-attention layers in the multimodal network.

A.11 Individual expert performance

The performance of the three experts varies across the three datasets. This manifests in the distribu-

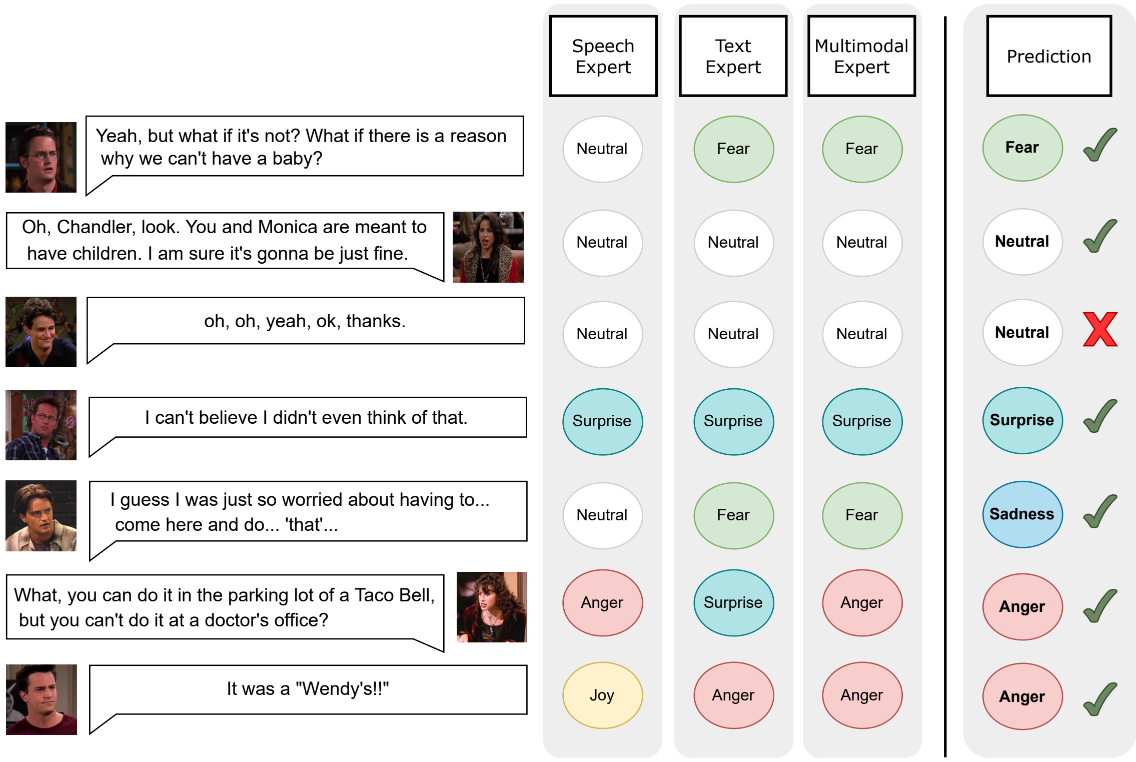


Figure 9: A case study from MELD where there are emotion shifts throughout the conversation. Out of the 7 utterances in the conversation, MiSTER-E correctly predicts 6.

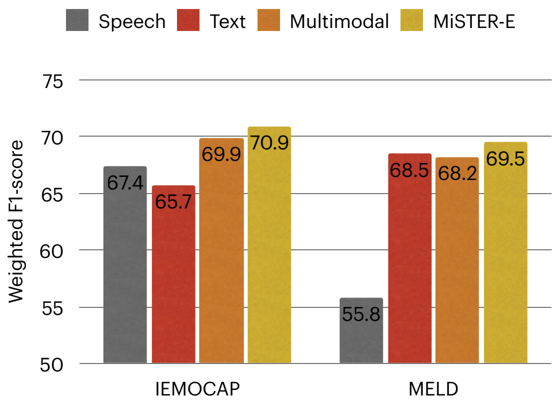


Figure 10: Performance of the experts on the two datasets.

tion of the weights assigned by the gating strategy (Sec. 5). We further report the performance of the individual experts in Fig. 10. For MELD, the performance of the audio expert is the lowest (55.8%). The multimodal expert is seen to slightly under-perform as compared to the text modality (by 0.3%) as it tries to align two modalities with significantly different capabilities. In the case of IEMOCAP, since both the modalities perform similarly, the multimodal expert is able to fuse the two representations to give the best performance. In this case, it outperforms the best unimodal expert (speech) by 2.5%.

Interestingly, we note that MiSTER-E outperforms the best expert (by 1% for both datasets). This indicates the effectiveness of our gating strategy and its ability to give more importance to the more dominant modality for improved ERC.

A.12 The Expert Consistency Loss

The value of α (Eq. 16) decides on how much consistency we desire in the output of the three experts. Since for IEMOCAP, the speech and text modalities perform similarly, we use $\alpha = 0.1$ in this case. This improves the performance of MiSTER-E on IEMOCAP from 70.4% ($\alpha = 0$) to 70.9%. Further increase in the value of α hurts the performance of our proposed method.

MELD, on the other hand, is a dataset where the textual modality far outperforms the speech modality. Hence, enforcing a strong consistency regularization hurts the model performance significantly. Thus we use a small value of $\alpha = 1e-3$ in the case of MELD, and this leads to a 0.1% increase in the overall performance.

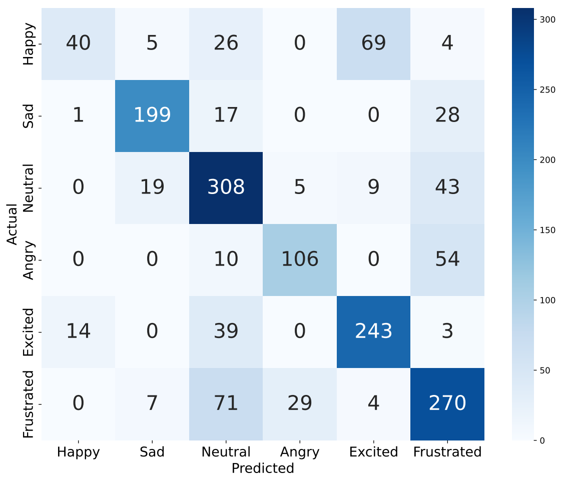


Figure 11: Confusion matrix for the IEMOCAP dataset.

A.13 Class-wise Performance of MiSTER-E

The class-wise performance of our proposed method for the two datasets is shown in Table 6. We note that for IEMOCAP the worst performing class is "happy" - likely because it is the least frequent class and it is most confused with the "excited" category (Fig. 11). Similarly, for MELD, the model performs most poorly on the "fear" category - again the least frequent class. However, we note that for both datasets, the model performs quite well on most emotion classes, thereby leading to an improved overall performance.

IEMOCAP		MELD	
Category	**F1**	**Category**	**F1**
Angry	68.4%	Angry	62.5%
Excited	77.9%	Disgust	42.9%
Frustrated	69.0%	Fear	30.6%
Happy	40.2%	Joy	65.4%
Neutral	80.2%	Neutral	81.3%
Sad	83.8%	Sad	50.3%
-	-	Surprise	61.5%

Table 6: Class wise performance of MiSTER-E on the two datasets. F1 stands for F1-score

A.14 Case Study

We provide a case study from the MELD dataset in Fig. 9. A number of key points that we wish to highlight from this example are as follows:

- MiSTER-E does not fail when there are emotion shifts. E.g., the conversation has 6 emotions in 7 utterances. Although, the model incorrectly predicts the third utterance to be neutral (instead of joy), it is able to predict the emotion shift from surprise to sadness to anger.
- Another interesting point is the output of the fifth utterance. None of the experts predict sadness, yet the MoE strategy correctly marks the utterance as sad. This indicates the utility of the MoE gating strategy in our proposed method and differentiates it from static ensembling techniques.

A.15 License

SALMONN, the Qwen models and Gemma are distributed under Apache License 2.0, while LLaMA is distributed under the LLaMA license [11]. The IEMOCAP dataset is available for use for academic purposes, while MELD is distributed under the GNU General Public License v3.0.

[11] https://www.llama.com/llama3/license/

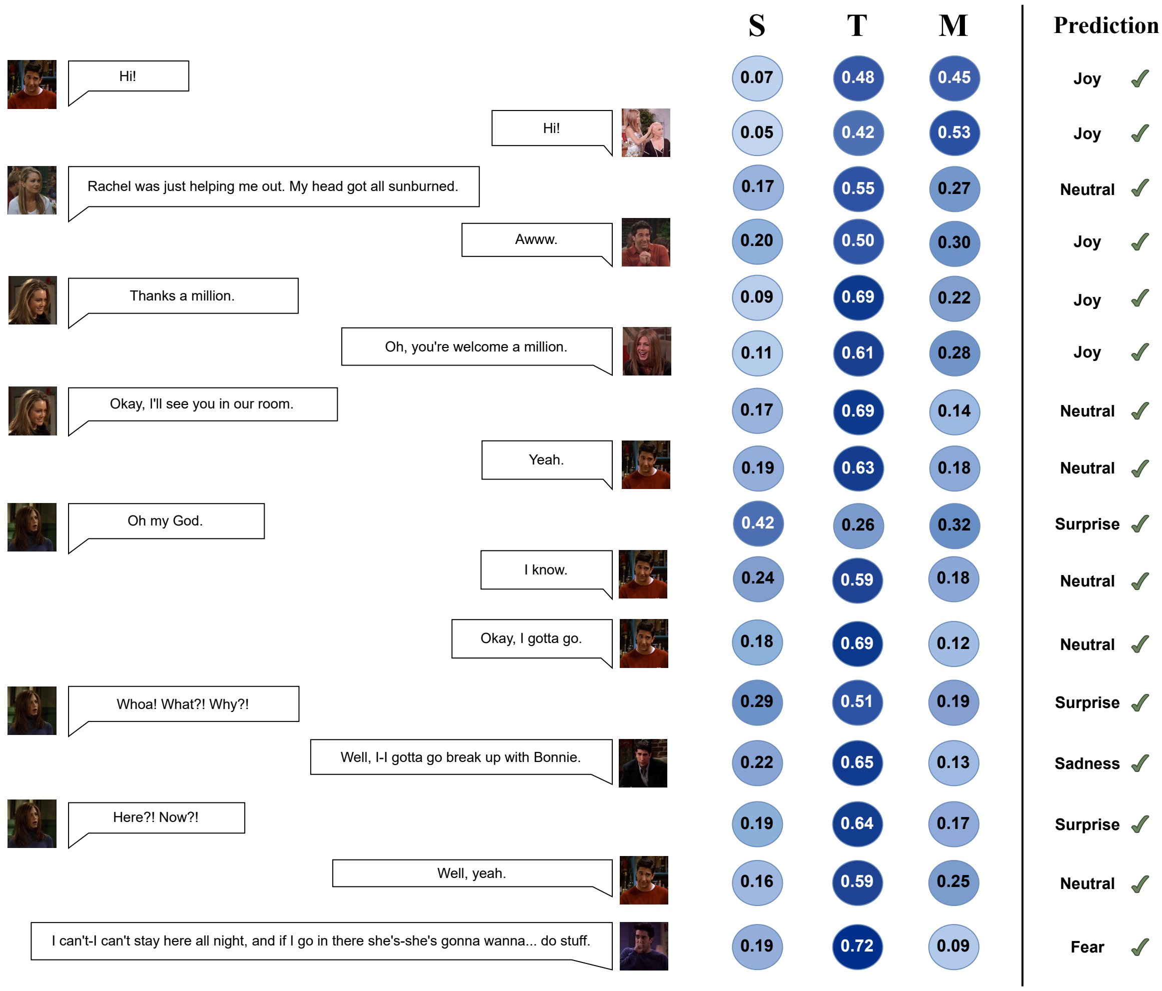


Figure 12: A case study from MELD where there are 16 utterances in the conversation. The weights assigned to the different experts are shown. In most cases, text is assigned the highest weight (except utterance 2 and 9. S: Speech Expert, T: Text Expert, M: Multimodal Expert. The predictions of MiSTER-E are also shown. The model correctly predicts all the utterances in the conversation, inspite of the emotion shifts present.