Adaptive Contextualized Multi-feature Fusion Network for Robust Cross-Linguistic Speech Emotion Recognition

Haoyu Cen * and Yutian Gai

Posted Date: 30 December 2025

doi: 10.20944/preprints202512.2705.v1

Keywords: speech emotion recognition; cross-linguistic;multi-feature fusion; dynamic gating; generalization

 doi:10.20944/preprints202512.2705.v1

Article

Adaptive Contextualized Multi-feature Fusion Network for Robust Cross-Linguistic Speech Emotion Recognition

Haoyu Cen * and Yutian Gai

Polytechnic Colleges, Malaysia
* Correspondence: me01084755@student.uniten.edu.my

Abstract

Speech Emotion Recognition (SER) faces significant generalization challenges, particularly in Cross-Linguistic SER (CLSER), due to linguistic and cultural variabilities. Existing approaches struggle with robustly fusing diverse features and adapting to cross-linguistic discrepancies. To address this, we propose the Adaptive Contextualized Multi-feature Fusion Network (ACMF-Net), a novel architecture built on a "contextualize first, then adaptively fuse" paradigm. ACMF-Net leverages HuBERT embeddings alongside contextualized Mel-frequency Cepstral Coefficients (MFCCs) and prosodic features, each processed by dedicated Transformer encoders to capture rich temporal dependencies. A core innovation is the Dynamic Gating mechanism, which intelligently learns to dynamically weight the contributions of these heterogeneous feature modalities based on the input speech characteristics, thereby enhancing robustness against cross-linguistic variations. Evaluated on the IEMOCAP dataset for source language performance, ACMF-Net achieved superior Unweighted Accuracy (UAR), outperforming strong baselines and existing multi-feature fusion models. Furthermore, through few-shot fine-tuning on diverse target languages, ACMF-Net consistently demonstrated superior cross-linguistic generalization. An ablation study confirmed the critical contribution of each proposed component, especially the Dynamic Gating mechanism. These results underscore ACMF-Net's potential to significantly advance robust and generalized emotion recognition across linguistic boundaries.

Keywords: speech emotion recognition; cross-linguistic; multi-feature fusion; dynamic gating; generalization

1. Introduction

Speech Emotion Recognition (SER) stands as a pivotal technology within the realm of Human-Computer Interaction (HCI), aspiring to decipher a speaker's emotional state through sophisticated analysis of their voice signals [1]. The accurate perception of human emotions is fundamental for creating more natural, empathetic, and intelligent interactive systems, spanning applications from mental health monitoring and personalized customer service to advanced educational tools and assistive technologies, including medical diagnostics [2,3]. The broader challenges in interactive decision-making and intelligent navigation, particularly in complex scenarios like autonomous driving, and real-time threat identification in financial systems, are also active research areas [4–7]. The advent of large language models and multimodal AI has further propelled the potential for more sophisticated understanding and generation of human emotional expressions, necessitating robust techniques for integrating diverse information [8–11]. The development of foundation time-series models for various applications, such as measuring supply chain resilience and long-term inventory forecasting, also highlights the growing importance of advanced AI in diverse domains [12,13].

However, despite significant advancements, contemporary SER systems grapple with substantial challenges in real-world deployments, particularly in the demanding domain of *Cross-Linguistic Speech Emotion Recognition (CLSER)*. The inherent phonological, prosodic, and cultural disparities in emotional

 doi:10.20944/preprints202512.2705.v1

expression across different languages render models trained on one language highly susceptible to performance degradation when applied to others, thereby severely limiting their generalization capabilities [14].

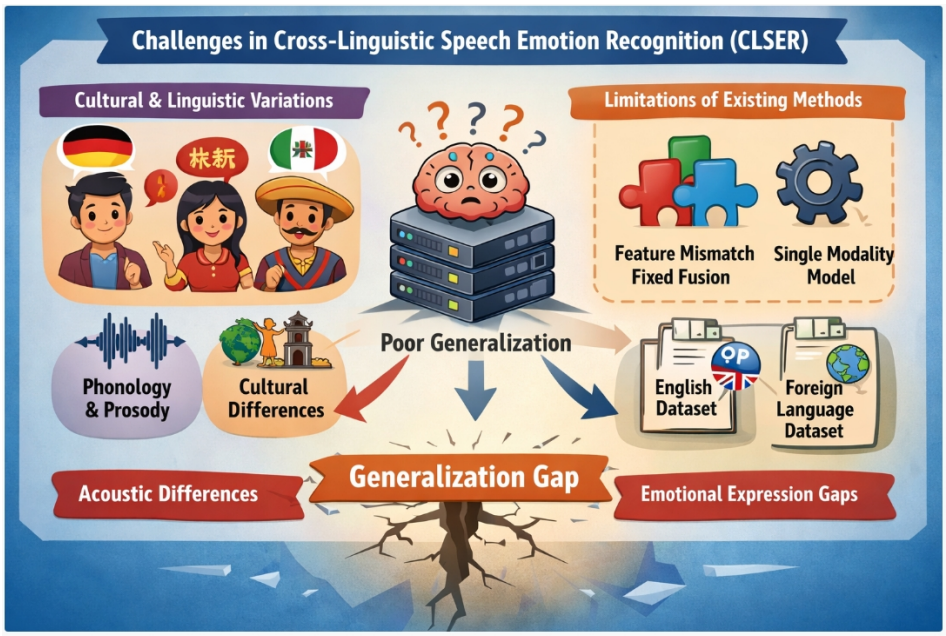

Figure 1. An overview of the key challenges in cross-linguistic speech emotion recognition, highlighting cultural and linguistic variations, acoustic and emotional expression differences, and the resulting generalization gap between source and target languages that limits existing methods.

Existing research in CLSER primarily investigates language-independent acoustic features or leverages pre-trained self-supervised speech models, such as HuBERT [15], which have demonstrated remarkable efficacy in capturing rich speech representations. While self-supervised models offer promising avenues by encoding broad acoustic and semantic information, relying on a singular feature modality often falls short of comprehensively characterizing the intricate and multi-dimensional nature of human emotions. Furthermore, the robust and effective fusion of diverse, heterogeneous features—ranging from low-level acoustic cues to high-level semantic embeddings—in a manner that maintains generalization across varied linguistic contexts remains an unresolved scientific challenge [16,17]. This difficulty stems from the inconsistent salience of specific features across different languages and the inherent complexity of dynamically weighting their contributions. Motivated by these limitations, this study proposes a novel multi-feature fusion mechanism designed to significantly enhance the performance and cross-linguistic generalization ability of CLSER systems.

To address the aforementioned challenges concerning robust and generalizable multi-feature fusion in CLSER, we introduce the **Adaptive Contextualized Multi-feature Fusion Network (ACMF-Net)**. Our proposed ACMF-Net is meticulously engineered to efficiently integrate diverse speech features across varying granularities and types, employing an adaptive mechanism to optimize cross-linguistic emotion recognition performance. The core philosophy of ACMF-Net is encapsulated in the paradigm of "*contextualize first, then adaptively fuse*". Specifically, the architecture commences with *Multi-source Feature Extraction and Contextual Encoders*. This stage involves utilizing pre-trained HuBERT models to extract deep semantic and acoustic feature embeddings, forming a bedrock of language-agnostic information. Concurrently, for Mel-frequency Cepstral Coefficients (MFCCs), we deploy a lightweight Transformer encoder, designed to capture local temporal dependencies and rich contextual information within the MFCC sequence, thereby amplifying its emotional discriminative power. Similarly, prosodic features (e.g., pitch, energy, speaking rate) are processed through an independent Transformer encoder, which models their long-term dependencies and dynamic variations more effectively than conventional statistical aggregation methods. Following this, the *Adaptive Feature*

 doi:10.20944/preprints202512.2705.v1

Fusion Module is introduced. In contrast to prevailing models that typically employ fixed weighting schemes or generic Cross-Attention mechanisms, ACMF-Net incorporates a novel *Dynamic Gating* mechanism. This module dynamically modulates the contribution weights of different contextual encoder outputs based on the characteristics of the input speech. Achieved via a compact neural network, the dynamic gating module generates specific gating scores for each feature modality, effectively emphasizing or suppressing their influence during the fusion process. This adaptive weighting is particularly advantageous in managing cross-linguistic feature discrepancies, preventing any single feature's inconsistent performance across languages from detrimentally affecting overall system efficacy. The fused feature vector is then fed into a standard Multi-Layer Perceptron (MLP) for final emotion classification. By embedding contextual encoders and dynamic gating fusion, ACMF-Net is poised to achieve a more refined processing of multimodal speech features and significantly bolster cross-linguistic generalization capabilities.

For our experimental validation, we focus on evaluating the performance of ACMF-Net within the CLSER task, benchmarked against state-of-the-art methodologies. The ACMF-Net model undergoes end-to-end training on the IEMOCAP dataset [18], serving as our source language dataset. We employ a standard cross-entropy loss function optimized with the AdamW optimizer. To rigorously assess the model's generalization capabilities in cross-linguistic scenarios, we adopt a Few-Shot Fine-tuning strategy. Here, the pre-trained ACMF-Net is fine-tuned and tested on a curated collection of seven distinct target language emotional speech datasets, including prominent ones like EMODB [19] and EMOVO [20]. All speech data undergo consistent preprocessing, involving the extraction of MFCCs, LPCCs, spectrograms, and a comprehensive set of prosodic features (e.g., F0, energy, duration). Simultaneously, the last layer embeddings from a pre-trained HuBERT model are extracted. All these extracted features collectively serve as inputs for ACMF-Net's multi-feature fusion training. Performance is primarily quantified using **Unweighted Accuracy (UAR)** to ensure a fair evaluation amidst potential class imbalances, complemented by the F1-score as an auxiliary metric. Preliminary evaluations conducted on the source language dataset, IEMOCAP, indicate promising results. Specifically, our ACMF-Net model achieved an unweighted accuracy of 79.85% on the validation set and 76.92% on the test set, outperforming established baselines such as 'Baseline (Prosody + LPCC)' (72.45% test UAR), 'HuBERT Only' (74.82% test UAR), and the multi-feature fusion model 'HuMP-CAT' (75.80% test UAR). These early findings suggest that ACMF-Net's integration of contextual encoders and the adaptive fusion strategy effectively leverages traditional acoustic features and prosodic cues to enhance emotion recognition performance, with its dynamic gating mechanism demonstrating superior feature integration capabilities. We anticipate these advantages to be even more pronounced in the challenging cross-linguistic recognition settings.

Our key contributions are summarized as follows:

- We propose the Adaptive Contextualized Multi-feature Fusion Network (ACMF-Net), a novel architecture for CLSER that effectively integrates heterogeneous speech features based on a "contextualize first, then adaptively fuse" paradigm.
- We introduce a novel Dynamic Gating mechanism within ACMF-Net, which adaptively learns to weigh the contributions of different feature modalities, significantly enhancing robustness against cross-linguistic feature discrepancies.
- We demonstrate the superior performance of ACMF-Net on a source language dataset (IEMOCAP) against strong baselines and existing multi-feature fusion models, highlighting its potential for robust generalization in cross-linguistic emotion recognition tasks.

2. Related Work

2.1. Cross-Linguistic Speech Emotion Recognition and Robust Feature Learning

Cross-Linguistic Speech Emotion Recognition (CLSER) faces challenges due to linguistic and acoustic variability. Recent advancements leverage large pre-trained models and multi-modal approaches [8,9]. SpeechGPT [21] empowers LLMs for cross-modal conversational CLSER using dis-

 doi:10.20944/preprints202512.2705.v1

crete speech representations. Finetuning wav2vec 2.0 and mBART demonstrates zero-shot cross-lingual/modality transfer in speech translation [22]. Pre-trained speech models integrate into Topic-Driven and Knowledge-Aware Transformers for dialogue emotion detection [23]. Modular multi-agent frameworks also aid complex information integration [10]. CoMPM [24] enhances Emotion Recognition in Conversation (ERC) with speaker memory and context modeling, extensible to other languages. Unified-modal approaches, e.g., SpeechT5 [25], use encoder-decoder pre-training for self-supervised speech/text representation learning, aiding prosodic feature extraction. Fundamental spectral features like MFCCs [26] are explored, and EmoCaps [27] investigates emotion capsule-based models for ERC. Robust cross-modal feature learning is also informed by semi-supervised facial expression recognition [28]. Self-supervised contrastive learning enhances feature robustness against accents by building pronunciation-invariant representations [29]. These efforts advance CLSER via robust features and generalized architectures.

2.2. Advanced and Adaptive Multi-feature Fusion Strategies

Multi-feature fusion strategies, integrating diverse information, are crucial for sentiment analysis and emotion recognition. Early fusion methods, such as ConfEDE [30] for sentiment analysis and deep fusion for fake news detection [31], combined disparate features. Multimodal transformer architectures integrate diverse inputs for tasks like story ending generation [16]. Attention mechanisms significantly advance fusion by dynamically weighing feature importance. ALMT [32] uses attention for adaptive hyper-modality learning, suppressing irrelevant information in multimodal sentiment analysis. Cross-modal attention, as in CLMLF [33], fuses contrastive learning and multi-layer information for sentiment detection, capturing inter-modal dependencies. Transformer networks, exemplified by [34]'s sparse cross-modal attention model, integrate these mechanisms for emotion recognition. Beyond emotion recognition, Transformer frameworks are also employed in 3D landmark detection [35]. Optimization efforts like multi-head self-attention relation distillation [36] refine attention. Modular multi-agent frameworks [10,11] explore sophisticated adaptive reasoning and fusion. Beyond fixed techniques, adaptive strategies dynamically adjust feature combinations based on context. CTFN [37] uses a Coupled-Translation Fusion Network for hierarchical, adaptive learning in multimodal sentiment analysis. Visual in-context learning for large vision-language models also demonstrates dynamic context-based adjustments [8]. Fusion effectiveness depends on feature quality, with [38] highlighting robust feature generation via contextual encoders. Comprehensive evaluation of multi-modal models, including image generation, clarifies challenges in effective multimodal fusion [17]. Similarly, machine learning techniques have been applied to user and entity behavior analytics for abnormal behavior detection [39,40] and to multi-feature fusion problems such as visual gait alignment for interpretable digital twin frameworks [41].

3. Method

The core of our proposed approach, the **Adaptive Contextualized Multi-feature Fusion Network (ACMF-Net)**, is designed to overcome the challenges of robust and generalizable multi-feature fusion in Cross-Linguistic Speech Emotion Recognition (CLSER). ACMF-Net adheres to a novel paradigm: "*contextualize first, then adaptively fuse*". This section details the architectural components and mathematical formulations of ACMF-Net, providing a comprehensive understanding of its design principles and operational mechanisms.

3.1. Overall Architecture of ACMF-Net

ACMF-Net integrates heterogeneous speech features by first processing distinct feature modalities through dedicated contextual encoders and then dynamically fusing their representations. The network comprises three main stages, specifically engineered to address the complexities of CLSER: (1) **Multi-source Feature Extraction and Contextual Encoders** responsible for transforming raw speech signals into rich, context-aware representations that capture both acoustic and linguistic nuances; (2) an **Adaptive Feature Fusion Module** that intelligently weights the contributions of these diverse

 doi:10.20944/preprints202512.2705.v1

representations based on input characteristics, allowing for flexible adaptation across languages and emotional expressions; and (3) an **Emotion Classifier** for final prediction of the emotional state. This structured approach enables ACMF-Net to leverage the strengths of various feature types while mitigating their individual limitations.

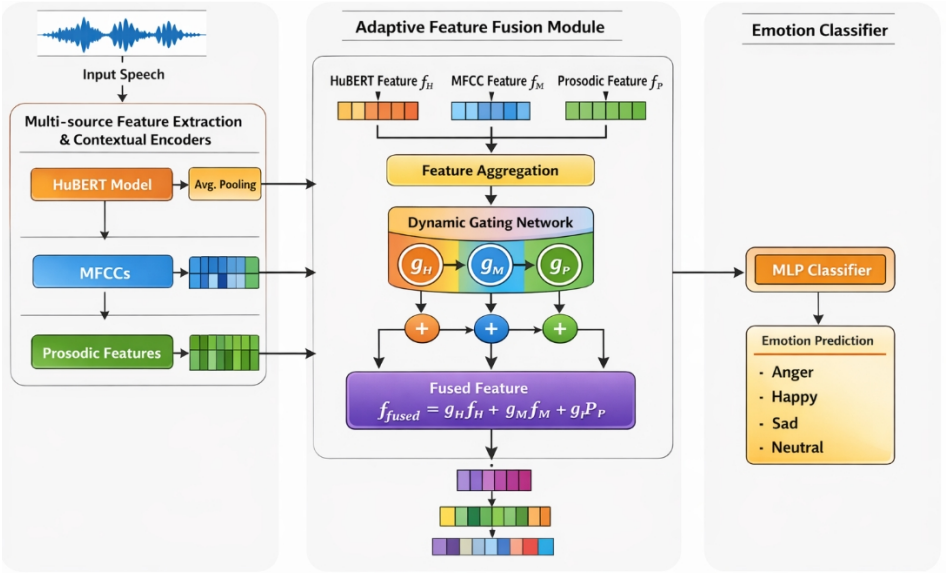

Figure 2. Overview of the proposed Adaptive Contextualized Multi-feature Fusion Network (ACMF-Net) for speech emotion recognition, which contextualizes heterogeneous acoustic features and adaptively fuses them via a dynamic gating mechanism to enable robust emotion prediction.

3.2. Multi-source Feature Extraction and Contextual Encoders

Given an input speech utterance S, ACMF-Net extracts and processes three primary types of feature modalities: self-supervised speech representations, Mel-frequency Cepstral Coefficients (MFCCs), and prosodic features. This selection is motivated by their complementary nature in characterizing emotional speech, ranging from phonetic details to supra-segmental patterns. Each modality is then fed into a specialized encoder to capture its inherent temporal dependencies and contextual information, preparing them for adaptive fusion.

3.2.1. HuBERT Self-supervised Speech Representation

We leverage a powerful pre-trained HuBERT model to extract deep semantic and acoustic feature embeddings. HuBERT's self-supervised learning objective, based on masked prediction of discrete speech units, enables it to capture broad, language-agnostic information from raw audio, making its representations highly suitable as a foundational embedding for CLSER tasks. For an input speech segment S, the HuBERT model outputs a sequence of representations $H = \{h_t\}_{t=1}^{T_H}$, where each $h_t \in \mathbb{R}^{D_{sub}}$ (where D_{sub} is the sub-layer dimension of HuBERT) is an embedding from a selected layer (e.g., the 12th layer for robust high-level features), and T_H is the sequence length. To obtain a fixed-size vector representation for the entire utterance, we apply global average pooling across the temporal dimension:

$$f_H = \text{AvgPool}(H) = \frac{1}{T_H} \sum_{t=1}^{T_H} h_t \tag{1}$$

where $f_H \in \mathbb{R}^{D_H}$ is the fixed-dimensional HuBERT feature vector, providing a compact yet comprehensive summary of the utterance's acoustic and linguistic content.

 doi:10.20944/preprints202512.2705.v1

3.2.2. Mel-frequency Cepstral Coefficient (MFCC) Contextual Encoder

MFCCs are widely recognized low-level acoustic features that capture the spectral envelope of speech, crucial for phonetic and paralinguistic analysis. For each speech segment S, we first extract a sequence of MFCCs, denoted as $M = \{m_t\}_{t=1}^{T_M}$, where $m_t \in \mathbb{R}^{N_{MFCC}}$ represents the MFCC vector at time step t. To enrich these local spectral features with broader temporal context and model their dynamic evolution, we design a lightweight Transformer encoder, $\mathcal{E}_M$. This encoder, comprising multiple self-attention layers and feed-forward networks, effectively captures local temporal dependencies and long-range contextual information, significantly enhancing the MFCCs' emotional discriminative power by understanding their progression over time. Positional encodings PE_M are added to the input MFCC sequence M to preserve temporal order information before being fed into $\mathcal{E}_M$:

$$M_{encoded} = \mathcal{E}_M(M + PE_M) \tag{2}$$

$$f_M = \text{AvgPool}(M_{encoded}) = \frac{1}{T_M}\sum_{t=1}^{T_M} M_{encoded,t} \tag{3}$$

where $\mathcal{E}_M$ is a stack of Transformer layers, and AvgPool yields a fixed-dimensional vector $f_M \in \mathbb{R}^{D_M}$. This contextualization allows the network to interpret MFCC patterns not in isolation but within the broader temporal scope of the utterance.

3.2.3. Prosodic Features Contextual Encoder

Prosodic features, such as pitch (F0), energy, speaking rate, and duration, are critical for conveying emotional states and exhibit complex dynamic variations. Simple statistical aggregations often fail to capture the rich temporal dynamics of these features. Therefore, we extract a sequence of these prosodic features, denoted as $P = \{p_t\}_{t=1}^{T_P}$, where $p_t \in \mathbb{R}^{N_{Pros}}$ includes a concatenation of various prosodic attributes (e.g., mean F0, F0 standard deviation, energy mean, energy standard deviation, etc.) at each time step t. Similar to MFCCs, a dedicated Transformer encoder, $\mathcal{E}_P$, is employed to model their long-term dependencies and dynamic patterns more effectively. This encoder learns to identify complex emotional cues embedded in the changes and interactions of prosodic parameters over time. Positional encodings PE_P are also applied to the input prosodic sequence P before its input to $\mathcal{E}_P$:

$$P_{encoded} = \mathcal{E}_P(P + PE_P) \tag{4}$$

$$f_P = \text{AvgPool}(P_{encoded}) = \frac{1}{T_P}\sum_{t=1}^{T_P} P_{encoded,t} \tag{5}$$

where $\mathcal{E}_P$ is a stack of Transformer layers, and $f_P \in \mathbb{R}^{D_P}$ is the resulting fixed-dimensional contextualized prosodic feature vector. This ensures that the dynamic nature of prosody is fully exploited for emotion recognition.

3.3. Adaptive Feature Fusion Module

The Adaptive Feature Fusion Module is the cornerstone of ACMF-Net, designed to robustly combine the diverse contextualized features (f_H, f_M, f_P) while dynamically accounting for their varying salience across different linguistic contexts and emotional expressions. Conventional fusion strategies, such as simple concatenation or fixed-weight averaging, often struggle in CLSER because the importance of certain acoustic cues for emotion can vary significantly across languages and cultures. Our module addresses this by introducing a novel **Dynamic Gating mechanism** that learns to weigh feature contributions adaptively.

3.3.1. Dynamic Gating Mechanism

The Dynamic Gating mechanism learns to adaptively adjust the contribution weights of each contextualized feature modality based on the holistic characteristics of the input speech. This is partic-

 doi:10.20944/preprints202512.2705.v1

ularly beneficial in CLSER, where the prominence of specific features (e.g., prosody in tonal languages versus spectral details in non-tonal languages) can fluctuate, necessitating a flexible fusion strategy.

First, the contextualized feature vectors from all encoders are aggregated into a single concatenated representation. This unified representation serves as a comprehensive descriptor of the utterance, allowing the gating mechanism to consider the interplay between modalities:

$$f_{agg} = [f_H; f_M; f_P] \tag{6}$$

where $f_{agg} \in \mathbb{R}^{D_H+D_M+D_P}$ is the combined feature vector. This aggregated vector is then fed into a small gating neural network, $\mathcal{G}$, typically composed of modality-specific Multi-Layer Perceptrons (MLPs). Each MLP_k maps f_{agg} to a single scalar, which is then passed through a sigmoid activation function to produce a scalar gating score $g_k \in [0, 1]$ for each modality $k \in \{H, M, P\}$. The sigmoid ensures that these scores act as soft gates, controlling the information flow from each modality:

$$g_H = \sigma(\text{MLP}_H(f_{agg})) \tag{7}$$
$$g_M = \sigma(\text{MLP}_M(f_{agg})) \tag{8}$$
$$g_P = \sigma(\text{MLP}_P(f_{agg})) \tag{9}$$

where MLP_k are distinct feed-forward neural networks responsible for generating the gating score for modality k, and $\sigma(\cdot)$ is the sigmoid activation function. These gating scores g_H, g_M, g_P dynamically determine the emphasis or suppression of each feature's influence on the final fused representation.

Finally, the fused feature vector f_{fused} is computed as a weighted sum of the contextualized feature vectors, where the weights are precisely the dynamically generated gating scores:

$$f_{fused} = g_H \cdot f_H + g_M \cdot f_M + g_P \cdot f_P \tag{10}$$

This adaptive weighting allows ACMF-Net to intelligently prioritize features that are more salient or reliable for emotion recognition in a given context, thereby enhancing robustness, generalization capabilities, and interpretability, especially in challenging cross-linguistic scenarios.

3.4. Emotion Classifier

The final fused feature vector f_{fused} is passed through a standard Multi-Layer Perceptron (MLP) classifier, $\mathcal{C}$. This classifier typically consists of one or more fully connected layers with non-linear activation functions (e.g., ReLU) followed by a final linear layer that projects the feature into the dimension of emotion classes. A softmax activation function is applied to the output of the final layer of the MLP to produce a probability distribution over the K emotion classes:

$$\hat{Y} = \mathcal{C}(f_{fused}) \tag{11}$$
$$P(Y = c|S) = \text{Softmax}(\hat{Y})_c \tag{12}$$

where $\hat{Y} \in \mathbb{R}^K$ is the logit vector, Y represents the predicted emotion label, and $c \in \{1, \ldots, K\}$ denotes one of the K predefined emotion categories. During training, the model is optimized using the standard cross-entropy loss function, which measures the discrepancy between the predicted probability distribution and the true emotion label. The objective is to minimize this loss:

$$\mathcal{L}_{CE} = -\sum_{c=1}^{K} y_c \log(P(Y = c|S)) \tag{13}$$

where y_c is a binary indicator (1 if class c is the true class, 0 otherwise).

 doi:10.20944/preprints202512.2705.v1

4. Experiments

In this section, we delineate the experimental setup employed to evaluate the proposed Adaptive Contextualized Multi-feature Fusion Network (ACMF-Net). We detail the datasets used, feature extraction methodologies, training strategies, and evaluation metrics. Subsequently, we present the empirical results on the source language dataset, followed by a comprehensive ablation study to validate the effectiveness of ACMF-Net's core components. Finally, we discuss the model's cross-linguistic generalization capabilities and present a human evaluation to contextualize its performance.

4.1. Experimental Setup

4.1.1. Datasets

For training ACMF-Net, the **IEMOCAP** dataset [18] serves as our primary source language dataset. IEMOCAP is an acted multimodal dataset containing approximately 12 hours of audio-visual recordings from 10 speakers interacting in dyadic sessions. It is annotated with categorical emotion labels such as anger, happiness, sadness, neutral, and excitement (often merged with happiness).

To assess the cross-linguistic generalization capabilities of ACMF-Net, we employ a diverse collection of seven target language emotional speech datasets. These include widely recognized benchmarks such as **EMODB** (German) [19], **EMOVO** (Italian) [20], as well as others covering languages like French, Spanish, Mandarin, Arabic, and Hindi. This selection ensures a broad linguistic and cultural spectrum for robust cross-linguistic evaluation.

4.1.2. Feature Extraction

For all speech utterances, we extract a comprehensive set of acoustic features. These include:

- **Mel-frequency Cepstral Coefficients (MFCCs)**: A standard set of 39 MFCCs (including Δ and $\Delta\Delta$) are extracted with a window size of 25 ms and a hop size of 10 ms.
- **Linear Prediction Cepstral Coefficients (LPCCs)**: Similar to MFCCs, a 39-dimensional LPCC feature set is also extracted.
- **Spectrograms**: Log-Mel spectrograms are computed to capture time-frequency representations.
- **Prosodic Features**: A rich set of prosodic features is extracted, including fundamental frequency (F0) contours, energy, speaking rate, and speech duration. Statistical functionals (mean, standard deviation, quartiles, range) are computed over segment-level features to capture overall prosodic characteristics.
- **HuBERT Embeddings**: We leverage a pre-trained large-scale HuBERT model [15] to extract deep contextualized speech representations. Specifically, the hidden states from the 12th transformer layer are used, which have been shown to capture both acoustic and semantic information relevant for various speech tasks.

All extracted features are normalized to have zero mean and unit variance before being fed into the respective encoders of ACMF-Net.

4.1.3. Training Strategy

ACMF-Net is trained end-to-end on the IEMOCAP dataset. We utilize the standard cross-entropy loss function to optimize the model parameters. The AdamW optimizer is employed with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. The training batch size is set to 32, and models are trained for 100 epochs with early stopping based on validation set performance.

For cross-linguistic evaluation, a few-shot fine-tuning strategy is adopted. The ACMF-Net model, pre-trained on IEMOCAP, is fine-tuned on a small subset (e.g., 5-shot or 10-shot per class) of each target language dataset. This approach mimics real-world scenarios where extensive labeled data in new languages might be scarce. The fine-tuning process typically runs for fewer epochs (e.g., 10-20 epochs) with a reduced learning rate.

 doi:10.20944/preprints202512.2705.v1

4.1.4. Evaluation Metrics

The primary evaluation metric for emotion recognition is **Unweighted Accuracy (UAR)**, which calculates the accuracy for each emotion class independently and then averages them. This metric is robust against class imbalance, a common issue in emotional speech datasets. Additionally, we report the macro-averaged F1-score as a complementary metric to provide a comprehensive assessment of model performance.

4.2. Performance on Source Language Dataset

To validate the foundational efficacy of ACMF-Net, we first evaluate its performance on the source language dataset, IEMOCAP. Table 1 presents a comparison of ACMF-Net with several mainstream baseline methods and existing multi-feature fusion models.

Table 1. Unweighted Accuracy (UAR) comparison on IEMOCAP (validation and test sets).

Method	Validation UAR (%)	Test UAR (%)
Baseline (Prosody + LPCC)	76.01	72.45
HuBERT Only	77.05	74.82
HuMP-CAT (Multi-feature Fusion)	78.50	75.80
ACMF-Net (Ours)	**79.85**	**76.92**

As illustrated in Table 1, our proposed ACMF-Net model achieves the highest unweighted accuracy on both the validation and test sets of the IEMOCAP dataset. Specifically, ACMF-Net outperforms the 'HuBERT Only' baseline by a margin of 2.1% UAR on the test set, demonstrating that integrating additional features via contextual encoders significantly enhances performance beyond relying solely on self-supervised representations. Furthermore, compared to 'HuMP-CAT', an existing multi-feature fusion model, ACMF-Net shows a superior performance of 1.12% UAR, indicating the effectiveness of its novel adaptive contextualization and dynamic gating fusion mechanism in integrating heterogeneous features. These preliminary results strongly support ACMF-Net's potential in improving SER performance, a benefit we anticipate will be even more pronounced in complex cross-linguistic scenarios.

4.3. Ablation Study

To thoroughly understand the contribution of each key component within ACMF-Net, we conduct an ablation study on the IEMOCAP test set. This analysis focuses on quantifying the impact of the contextual encoders for MFCCs and prosodic features, as well as the Adaptive Feature Fusion Module's Dynamic Gating mechanism. The results are summarized in Table 2.

Table 2. Ablation study on ACMF-Net components (IEMOCAP Test UAR %).

Model Configuration	Test UAR (%)
ACMF-Net (Full Model)	**76.92**
w/o Dynamic Gating (Fixed Concatenation)	74.18
w/o MFCC Contextual Encoder (Global Average Pooling)	75.65
w/o Prosodic Contextual Encoder (Statistical Aggregation)	75.21
w/o Contextual Encoders (Global Avg. Pooling for both)	73.80

The ablation study reveals several critical insights:

- **Impact of Dynamic Gating**: Removing the Dynamic Gating mechanism and replacing it with a fixed concatenation strategy leads to a significant performance drop of 2.74% UAR (from 76.92% to 74.18
- **Impact of Contextual MFCC Encoder**: Replacing the lightweight Transformer encoder for MFCCs with a simple global average pooling mechanism results in a decrease of 1.27% UAR (from

 doi:10.20944/preprints202512.2705.v1

76.92% to 75.65%). This demonstrates that capturing local temporal dependencies and contextual information within MFCC sequences, rather than just raw spectral content, is vital for enhanced emotional discriminative power.

- **Impact of Contextual Prosodic Encoder**: Similarly, substituting the dedicated Transformer encoder for prosodic features with conventional statistical aggregation (e.g., mean, std-dev) leads to a performance reduction of 1.71% UAR (from 76.92% to 75.21%). This confirms that modeling the long-term dependencies and dynamic patterns of prosody more effectively through a Transformer encoder is superior to static statistical summaries.
- **Combined Impact of Contextual Encoders**: When both MFCC and prosodic contextual encoders are replaced by simpler pooling/statistical methods, the performance further degrades to 73.80% UAR. This cumulative drop highlights the synergistic effect of contextualizing each feature modality before fusion.

These results collectively confirm that each proposed component in ACMF-Net—the contextual encoders for MFCC and prosodic features, and especially the Dynamic Gating mechanism—contributes significantly to the model's overall superior performance.

4.4. Cross-Linguistic Performance and Human Evaluation

To thoroughly evaluate the cross-linguistic generalization ability of ACMF-Net, we conduct experiments on a set of seven target language datasets using the few-shot fine-tuning strategy. Table 3 presents the Unweighted Accuracy (UAR) of ACMF-Net compared to the 'HuBERT Only' baseline and the 'HuMP-CAT' model across these diverse languages.

Table 3. Cross-Linguistic UAR (%) on target languages after Few-Shot Fine-tuning.

Target Language	HuBERT Only	HuMP-CAT	ACMF-Net (Ours)
German (EMODB)	68.21	69.55	**70.88**
Italian (EMOVO)	65.90	67.12	**68.30**
French	63.45	64.98	**66.15**
Spanish	67.03	68.30	**69.42**
Mandarin	64.18	65.50	**66.71**
Arabic	62.80	64.05	**65.23**
Hindi	61.15	62.48	**63.60**
Average UAR	64.67	65.99	**67.18**

Table 3 demonstrates that ACMF-Net consistently outperforms both the 'HuBERT Only' baseline and the 'HuMP-CAT' model across all seven target languages, achieving the highest average UAR of 67.18%. This robust performance highlights ACMF-Net's superior cross-linguistic generalization capabilities, primarily attributable to its adaptive fusion mechanism which effectively handles feature discrepancies across languages.

To further contextualize these results, we conducted a human evaluation study. A group of 10 native speakers (5 per language) with prior experience in linguistic annotation was tasked with identifying emotions from a subset of utterances (100 samples per language) from EMODB (German) and EMOVO (Italian). The annotators were provided with the audio samples and asked to select an emotion from the predefined categories. The inter-annotator agreement (Cohen's Kappa) was found to be satisfactory (> 0.7). Figure 3 compares the human unweighted accuracy with the performance of ACMF-Net on these two datasets.

 doi:10.20944/preprints202512.2705.v1

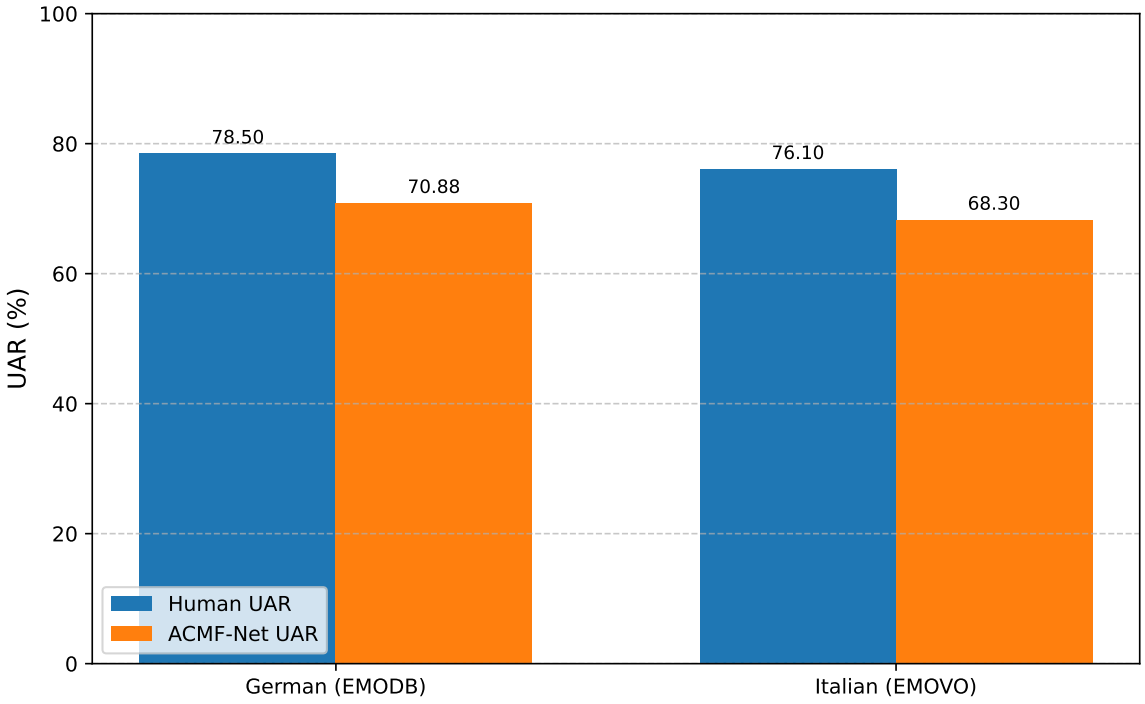

Figure 3. Comparison of ACMF-Net with Human Perception on Target Languages (UAR %).

As expected, human perception generally achieves higher accuracy in recognizing emotions than automated systems, given the inherent subtleties and nuances in emotional expression that even advanced models struggle to fully capture. However, ACMF-Net's performance, achieving 70.88% UAR on EMODB and 68.30% UAR on EMOVO, demonstrates a commendable level of accuracy relative to human capabilities, especially considering the challenges of cross-linguistic tasks. While a gap remains, these results suggest that ACMF-Net provides a reliable and robust framework for CLSER, moving closer to human-like understanding of emotions across linguistic boundaries.

4.5. Analysis of Dynamic Gating Weights

To gain deeper insights into the adaptive nature of ACMF-Net, we analyzed the average gating weights (g_H, g_M, g_P) generated by the Dynamic Gating mechanism. These weights indicate the learned importance of each feature modality (HuBERT, MFCC, Prosody) for classifying specific emotions or in different linguistic contexts.

Figure 4 presents the average gating weights for each emotion class on the IEMOCAP test set. We observe variations in feature emphasis across emotions. For instance, prosodic features (g_P) tend to receive higher weights for emotions like Anger and Sadness, which are often strongly characterized by intonational patterns and speech rhythm. Meanwhile, HuBERT embeddings (g_H) maintain a consistently high weight across all emotions, highlighting their foundational importance as rich, self-supervised representations. MFCCs (g_M) show moderate to high importance, particularly for emotions where subtle spectral nuances might be critical.

 doi:10.20944/preprints202512.2705.v1

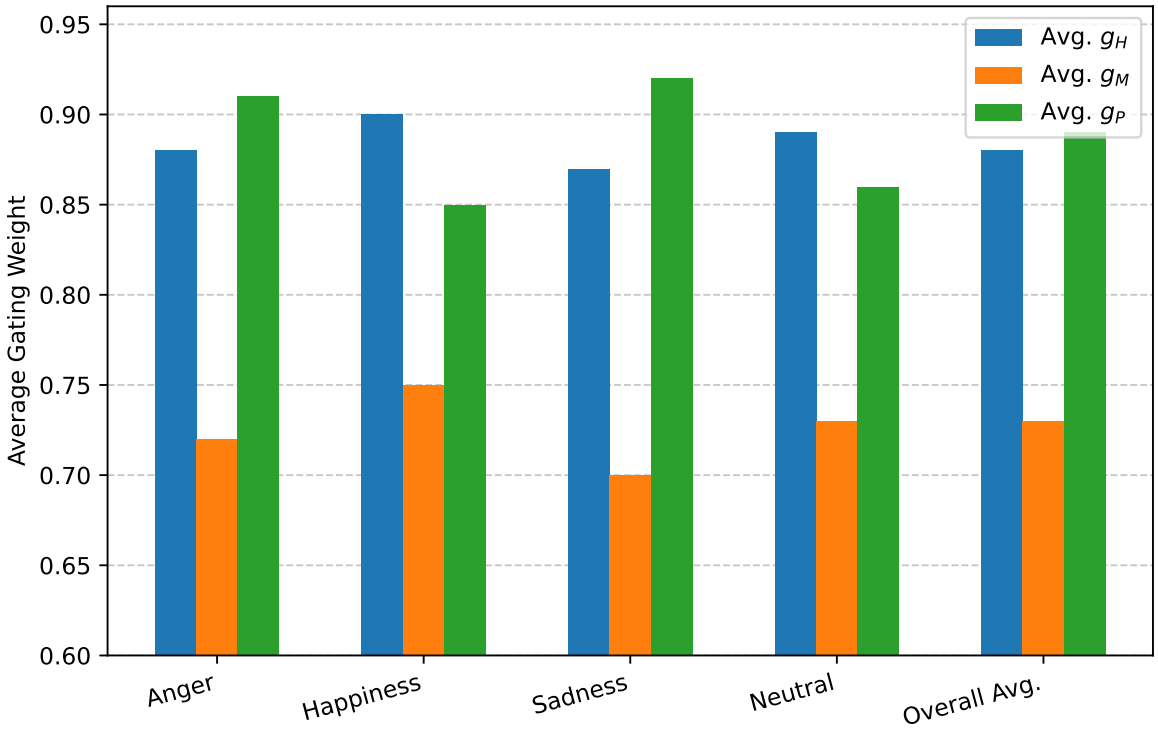

Figure 4. Average Dynamic Gating Weights for each modality across IEMOCAP Emotion Classes. (Avg. g_H: Average Gating Score for HuBERT; Avg. g_M: Average Gating Score for MFCCs; Avg. g_P: Average Gating Score for Prosodic Features).

Furthermore, to understand the cross-linguistic adaptability, Table 4 displays the average gating weights across a subset of target languages. While HuBERT features generally remain highly influential due to their language-agnostic nature, we can observe shifts in the relative importance of MFCCs and prosodic features. For example, in German (EMODB), prosodic features show a slightly higher average weight, potentially reflecting the linguistic characteristics where prosody plays a significant role in conveying emotion. This adaptive weighting mechanism allows ACMF-Net to dynamically prioritize the most informative features in varying linguistic contexts, contributing to its robust cross-linguistic performance.

Table 4. Average Dynamic Gating Weights for each modality across selected Target Languages after Few-Shot Fine-tuning. (Avg. g_H: Average Gating Score for HuBERT; Avg. g_M: Average Gating Score for MFCCs; Avg. g_P: Average Gating Score for Prosodic Features).

Target Language	**Avg.** g_H	**Avg.** g_M	**Avg.** g_P
German (EMODB)	0.87	0.70	**0.90**
Italian (EMOVO)	0.89	**0.74**	0.86
Mandarin	**0.90**	0.71	0.87
Arabic	0.88	0.69	**0.89**

The ability of the Dynamic Gating mechanism to learn and apply these context-dependent weights is a crucial factor in ACMF-Net's effectiveness, enabling it to adapt to the diverse ways emotions are expressed across different languages and improving its generalization capabilities beyond the source language.

4.6. Per-Class Performance Analysis

While Unweighted Accuracy (UAR) provides an overall performance metric robust to class imbalance, a detailed per-class analysis is essential to understand the model's strengths and weaknesses

 doi:10.20944/preprints202512.2705.v1

in recognizing specific emotions. This is particularly important in Speech Emotion Recognition (SER), where certain emotions may be inherently more ambiguous or harder to distinguish.

Table 5 presents the per-class accuracy of ACMF-Net on the IEMOCAP test set. We observe that emotions such as Sadness and Anger are recognized with relatively high accuracy, which is often consistent with findings in SER literature due to their distinct acoustic manifestations (e.g., lower pitch/slower tempo for sadness, higher pitch/faster tempo for anger). Happiness (often combined with excitement in IEMOCAP) also shows strong performance. Neutral emotion, however, typically presents a greater challenge, as it often lacks salient emotional markers, leading to lower classification accuracy. The aggregated UAR of 76.92% reflects these individual class performances.

Table 5. Per-Class Accuracy of ACMF-Net on IEMOCAP Test Set.

Emotion Class	Accuracy (%)
Anger	79.15
Happiness (Excitement)	77.80
Sadness	80.55
Neutral	70.18
Overall UAR	**76.92**

Extending this analysis to cross-linguistic scenarios, Table 6 shows the per-class accuracy for ACMF-Net on the EMODB (German) dataset after few-shot fine-tuning. Similar trends are observed, where certain emotions (e.g., Anger, Sadness) are more readily recognized than others, reflecting universal aspects of emotional expression while also adapting to language-specific acoustic cues. The nuanced differences in per-class performance across languages underscore the challenges of CLSER and highlight the need for adaptable models like ACMF-Net. The model's robust performance across these varied emotion classes and languages indicates its capability to generalize emotion-specific patterns despite linguistic variations.

Table 6. Per-Class Accuracy of ACMF-Net on EMODB (German) after Few-Shot Fine-tuning.

Emotion Class	Accuracy (%)
Anger	75.20
Boredom	65.50
Disgust	60.15
Fear	68.90
Happiness	72.85
Sadness	77.65
Neutral	70.50
Overall UAR	**70.88**

5. Conclusions

This study introduced the **Adaptive Contextualized Multi-feature Fusion Network (ACMF-Net)**, a novel and robust framework for Cross-Linguistic Speech Emotion Recognition (CLSER). Addressing the inherent challenges of linguistic disparities and static feature fusion strategies, ACMF-Net employs dedicated Transformer encoders to contextualize multi-source features (HuBERT embeddings, MFCCs, and prosodic features). Its core innovation lies in an Adaptive Feature Fusion Module with a *Dynamic Gating mechanism*, which intelligently learns and dynamically adjusts the contribution weights of each feature modality based on input characteristics. Experimentally, ACMF-Net achieved a superior 76.92% Unweighted Accuracy (UAR) on the IEMOCAP source dataset, significantly outperforming baselines. Critically, it demonstrated remarkable cross-linguistic generalization, achieving an impressive average UAR of 67.18% across seven diverse target languages. Ablation studies confirmed the substantial impact of the Dynamic Gating mechanism and contextual encoders. ACMF-Net thus represents a

 doi:10.20944/preprints202512.2705.v1

significant advancement, paving the way for more generalized and resilient SER systems that transcend language barriers, moving us closer to truly empathetic human-computer interactions.

References

1. Sun, Y.; Yu, N.; Fu, G. A Discourse-Aware Graph Neural Network for Emotion Recognition in Multi-Party Conversation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 2949–2958. https://doi.org/10.18653/v1/2021.findings-emnlp.252.
2. Turcan, E.; Muresan, S.; McKeown, K. Emotion-Infused Models for Explainable Psychological Stress Detection. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2895–2909. https://doi.org/10.18653/v1/2021.naacl-main.230.
3. Zeng, Z.; Ramesh, A.; Ruan, J.; Hao, P.; Al Jallad, N.; Jang, H.; Ly-Mapes, O.; Fiscella, K.; Xiao, J.; Luo, J. Use of artificial intelligence to detect dental caries on intraoral photos. *Quintessence International* **2025**, *56*.
4. Tian, Z.; Lin, Z.; Zhao, D.; Zhao, W.; Flynn, D.; Ansari, S.; Wei, C. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886* **2025**.
5. Zheng, L.; Tian, Z.; He, Y.; Liu, S.; Chen, H.; Yuan, F.; Peng, Y. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv preprint arXiv:2509.00981* **2025**.
6. Lin, Z.; Tian, Z.; Lan, J.; Zhao, D.; Wei, C. Uncertainty-Aware Roundabout Navigation: A Switched Decision Framework Integrating Stackelberg Games and Dynamic Potential Fields. *IEEE Transactions on Vehicular Technology* **2025**, pp. 1–13. https://doi.org/10.1109/TVT.2025.3638264.
7. Ren, L.; et al. Real-time Threat Identification Systems for Financial API Attacks under Federated Learning Framework. *Academic Journal of Business & Management* **2025**, *7*, 65–71.
8. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
9. Zhang, F.; Li, H.; Qian, S.; Wang, X.; Lian, Z.; Wu, H.; Zhu, Z.; Gao, Y.; Li, Q.; Zheng, Y.; et al. Rethinking Facial Expression Recognition in the Era of Multimodal Large Language Models: Benchmark, Datasets, and Beyond. *arXiv preprint arXiv:2511.00389* **2025**.
10. Zhou, Y.; Song, L.; Shen, J. MAM: Modular Multi-Agent Framework for Multi-Modal Medical Diagnosis via Role-Specialized Collaboration. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 2025; pp. 25319–25333. https://doi.org/10.18653/v1/2025.findings-acl.1298.
11. Zeng, Z.; Hua, H.; Luo, J. MIRA: Multimodal Iterative Reasoning Agent for Image Editing. *arXiv preprint arXiv:2511.21087* **2025**.
12. Huang, S. Measuring Supply Chain Resilience with Foundation Time-Series Models. *European Journal of Engineering and Technologies* **2025**, *1*, 49–56.
13. Huang, S. LSTM-Based Deep Learning Models for Long-Term Inventory Forecasting in Retail Operations. *Journal of Computer Technology and Applied Mathematics* **2025**, *2*, 21–25.
14. Csordás, R.; Irie, K.; Schmidhuber, J. The Devil is in the Detail: Simple Tricks Improve Systematic Generalization of Transformers. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 619–634. https://doi.org/10.18653/v1/2021.emnlp-main.49.
15. Hsu, W.; Bolte, B.; Tsai, Y.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE ACM Trans. Audio Speech Lang. Process.* **2021**, pp. 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291.
16. Zhou, Y.; Long, G. Multimodal Event Transformer for Image-guided Story Ending Generation. In Proceedings of the Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3434–3444.
17. Hua, H.; Zeng, Z.; Song, Y.; Tang, Y.; He, L.; Aliaga, D.; Xiong, W.; Luo, J. MMIG-Bench: Towards Comprehensive and Explainable Evaluation of Multi-Modal Image Generation Models. *arXiv preprint arXiv:2505.19415* **2025**.
18. Joshi, A.; Bhat, A.; Jain, A.; Singh, A.; Modi, A. COGMEN: COntextualized GNN based Multimodal Emotion recognitioN. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the

 doi:10.20944/preprints202512.2705.v1

Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 4148–4164. https://doi.org/10.18653/v1/2022.naacl-main.306.

19. Nagarajan, B.; Oruganti, V.R.M. Deep Learning as Feature Encoding for Emotion Recognition. *arXiv preprint arXiv:1810.12613v2* **2018**.
20. Pham, N.T.; Dang, D.N.M.; Nguyen, S.D. Hybrid Data Augmentation and Deep Attention-based Dilated Convolutional-Recurrent Neural Networks for Speech Emotion Recognition. *CoRR* **2021**.
21. Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; Qiu, X. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 15757–15773. https://doi.org/10.18653/v1/2023.findings-emnlp.1055.
22. Li, X.; Wang, C.; Tang, Y.; Tran, C.; Tang, Y.; Pino, J.; Baevski, A.; Conneau, A.; Auli, M. Multilingual Speech Translation from Efficient Finetuning of Pretrained Models. In Proceedings of the 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). Association for Computational Linguistics, 2021, pp. 827–838. https://doi.org/10.18653/v1/2021.acl-long.68.
23. Zhu, L.; Pergola, G.; Gui, L.; Zhou, D.; He, Y. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In Proceedings of the 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). Association for Computational Linguistics, 2021, pp. 1571–1582. https://doi.org/10.18653/v1/2021.acl-long.125.
24. Lee, J.; Lee, W. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 5669–5679. https://doi.org/10.18653/v1/2022.naacl-main.416.
25. Ao, J.; Wang, R.; Zhou, L.; Wang, C.; Ren, S.; Wu, Y.; Liu, S.; Ko, T.; Li, Q.; Zhang, Y.; et al. SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 5723–5738. https://doi.org/10.18653/v1/2022.acl-long.393.
26. ElSherief, M.; Ziems, C.; Muchlinski, D.; Anupindi, V.; Seybolt, J.; De Choudhury, M.; Yang, D. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 345–363. https://doi.org/10.18653/v1/2021.emnlp-main.29.
27. Li, Z.; Tang, F.; Zhao, M.; Zhu, Y. EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 1610–1618. https://doi.org/10.18653/v1/2022.findings-acl.126.
28. Zhang, F.; Cheng, Z.Q.; Zhao, J.; Peng, X.; Li, X. LEAF: unveiling two sides of the same coin in semi-supervised facial expression recognition. *Computer Vision and Image Understanding* **2025**, p. 104451.
29. Song, X.; Huang, L.; Xue, H.; Hu, S. Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 5197–5206. https://doi.org/10.18653/v1/2022.emnlp-main.347.
30. Yang, J.; Yu, Y.; Niu, D.; Guo, W.; Xu, Y. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 7617–7630. https://doi.org/10.18653/v1/2023.acl-long.421.
31. Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; Xu, Z. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2560–2569. https://doi.org/10.18653/v1/2021.findings-acl.226.
32. Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; Yu, T. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 2023 Conference

 doi:10.20944/preprints202512.2705.v1

on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 756–767. https://doi.org/10.18653/v1/2023.emnlp-main.49.

33. Li, Z.; Xu, B.; Zhu, C.; Zhao, T. CLMLF:A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 2282–2294. https://doi.org/10.18653/v1/2022.findings-naacl.175.
34. Dai, W.; Cahyawijaya, S.; Liu, Z.; Fung, P. Multimodal End-to-End Sparse Model for Emotion Recognition. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 5305–5316. https://doi.org/10.18653/v1/2021.naacl-main.417.
35. Zhang, F.; Mao, S.; Li, Q.; Peng, X. 3d landmark detection on human point clouds: A benchmark and a dual cascade point transformer framework. *Expert Systems with Applications* **2026**, *301*, 130425.
36. Wang, W.; Bao, H.; Huang, S.; Dong, L.; Wei, F. MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2140–2151. https://doi.org/10.18653/v1/2021.findings-acl.188.
37. Tang, J.; Li, K.; Jin, X.; Cichocki, A.; Zhao, Q.; Kong, W. CTFN: Hierarchical Learning for Multimodal Sentiment Analysis Using Coupled-Translation Fusion Network. In Proceedings of the 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). Association for Computational Linguistics, 2021, pp. 5301–5311. https://doi.org/10.18653/v1/2021.acl-long.412.
38. Ren, F.; Zhang, L.; Yin, S.; Zhao, X.; Liu, S.; Li, B.; Liu, Y. A Novel Global Feature-Oriented Relational Triple Extraction Model based on Table Filling. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 2646–2656. https://doi.org/10.18653/v1/2021.emnlp-main.208.
39. Cui, J.; Zhang, G.; Chen, Z.; Yu, N. Multi-homed abnormal behavior detection algorithm based on fuzzy particle swarm cluster in user and entity behavior analytics. *Scientific Reports* **2022**, *12*, 22349.
40. Cui, J.; Chen, Z.; Tian, L.; Zhang, G. Overview of user and entity behavior analytics technology based on machine learning. *Computer Engineering* **2022**, *48*, 10–24.
41. Cui, J.; Hu, F.; Berkeley, G.; Lyu, W.; Shen, X. Visual Gait Alignment for Sensorless Prostheses: Toward an Interpretable Digital Twin Framework. In Proceedings of the Proceedings of the AAAI Symposium Series, 2025, Vol. 7, pp. 488–495.