EAA: Emotion-Aware Audio Large Language Models with Dual Cross-Attention and Context-Aware Instruction Tuning

Hongfei Du[1], *Sidi Lu*[1], *Gang Zhou*[1], *Ye Gao*[1]

[1]Department of Computer Science, William & Mary, USA

hdu02@wm.edu, sidi@wm.edu, gzhou@wm.edu, ygao18@wm.edu

Abstract

Understanding speech emotion through artificial intelligence (AI) is crucial for human-computer interaction and mental health monitoring. While audio large language models (ALLMs) excel in speech comprehension, they face challenges in accurately integrating emotional signals from acoustic and semantic features. Moreover, emotions often span dialogues, making sole reliance on current audio insufficient for comprehensive understanding. To address these challenges, we propose a novel emotion-aware audio large language model (EAA). Specifically, we design a dual cross-attention mechanism to fuse acoustic and semantic information for a more comprehensive emotional representation. Furthermore, we use context-aware instruction tuning by incorporating the current and immediately preceding utterances as contextual information, enhancing task understanding and emotion recognition. Our experimental results show that EAA outperforms existing ALLMs on the MELD dataset, improving accuracy by 11.4%.

Index Terms: speech emotion recognition, audio large language models, instruction tuning

1. Introduction

Speech signals contain multi-level features, in which semantic information (e.g., speech content) and acoustic characteristics (e.g., pitch, timbre, and tone intensity) jointly shape emotional expression. Effectively capturing and integrating these features is essential for accurate emotion recognition, which plays a critical role in applications such as human-computer interaction, mental health monitoring, and customer service [1, 2, 3].

While large language models (LLMs) have demonstrated strong performance in text-based tasks [4], their ability to understand speech remains relatively underexplored. Recent advancements in audio large language models (ALLMs) have driven progress in speech processing [5, 6]. However, their ability to effectively perceive and integrate emotional cues remains limited due to insufficient feature integration and context modeling.

To be concrete, existing ALLMs **face two main challenges**. First, they normally have limitations in feature integration, as most models rely solely on a single audio encoder [5, 6], while others separately extract semantic and acoustic features and merely concatenate them [7], failing to capture the intricate interactions between between these two feature types. Second, emotions in dialogue evolve progressively within context, making it difficult to accurately perceive the overall emotional state from isolated audio segments. Most existing ALLMs do not effectively incorporate contextual information [8, 7], further limiting their ability to capture the dynamic nature of emotions.

To this end, this paper proposes EAA: emotion-aware audio LLMs with dual cross-attention to integrate semantic and acoustic features and context-aware instruction tuning for incorporating contextual information. This method can not only adaptively fuse semantic and acoustic features, but also make full use of dialogue context information, thus improving the accuracy of audio emotion recognition in ALLMs.

To effectively capture emotional information in speech, EAA use an acoustic encoder to extract acoustic features and a semantic encoder to process linguistic representations, with a dual cross-attention mechanism dynamically adjusting feature importance based on context to enhance integration. When emotional cues (e.g., high pitch for anger, slow speech for sadness) predominate, the model learns to assign greater weight to acoustic features; when explicit words (e.g., happy, angry) dominate, the model prioritizes linguistic features. Specifically, acoustic features serve as the query, while semantic features act as the key and value. Conversely, semantic features can also function as the query, with acoustic features as the key and value, forming a dual cross-attention mechanism. To maintain feature integrity while enhancing fusion, we concatenate the original acoustic and semantic features with the dual cross-attention outputs in the final fusion stage. This straightforward approach preserves unique feature characteristics while ensuring comprehensive integration.

In speech emotion recognition tasks, emotional expressions are highly context-dependent, making classification challenging when relying on a single audio instance. For example, the utterance *"What?"* can express anger, neutral, or surprise, depending on the preceding sentence, such as *"Did you just insult me?"*, or some shocking information. Without context, even combining semantic and acoustic features may be insufficient. To address this challenge, we propose a context-aware instruction tuning approach, which allows the model to consider both the current and preceding audio utterances. This technique enables the model to capture shifts in emotion more effectively, improving emotion state inference in context-dependent dialogue scenarios and ultimately enhancing the accuracy of speech emotion recognition systems.

The main contributions of this paper are as follows: (1) We propose a novel dual cross-attention mechanism that integrates acoustic and semantic features, capturing their interactions while preserving expressiveness to enhance audio representation. (2) We introduce context-aware instruction tuning that integrates the current and immediately preceding utterance as contextual cues, the model better captures emotion evolution, reduces ambiguity, and improves adaptability for speech emotion recognition tasks. (3) Comprehensive experiments demonstrated the superiority of our proposed methods, achieving an improvement of 11.4%.

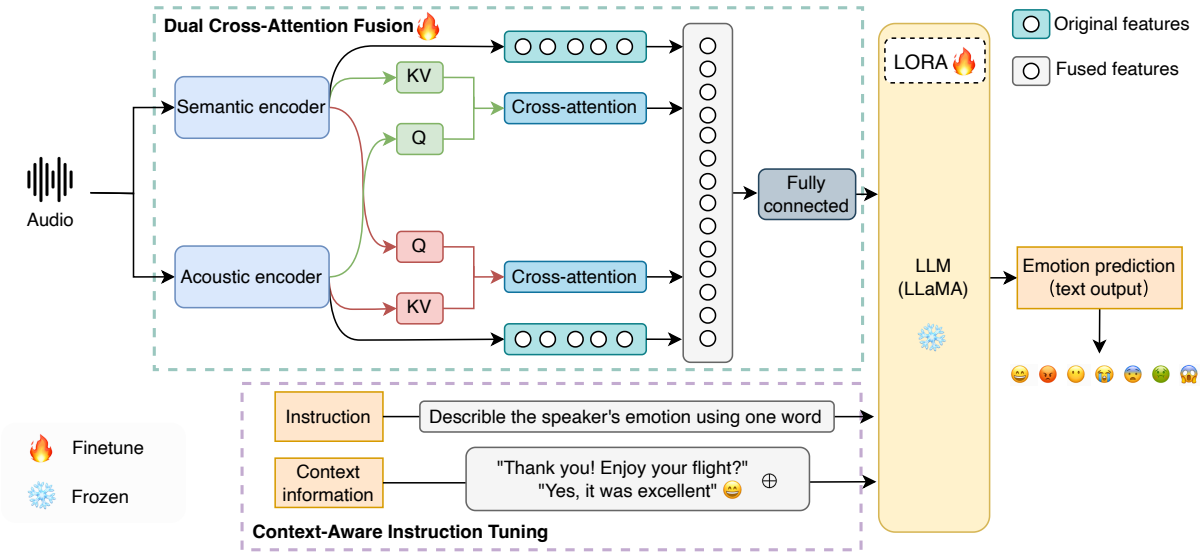

Figure 1: *The architecture overview of the proposed EAA.*

2. Related Works

Current research on audio large language models explores various encoder choices, utilizing different fusion strategies for multiple encoders. Pengi [5] is the first model to integrate audio encoders with LLMs, utilizing CLAP [9] as the audio encoder. MERaLiON-AudioLLM [10] employs MERaLiON-Whisper as its audio encoder and uses a two-layer MLP to align with text input. Qwen-Audio [6] and OSUM [11] also opt for Whisper as the audio encoder. These models rely on a single encoder for feature extraction. In contrast, SALMONN [8] uses both Whisper and BEATs encoders, using a fusion approach based on splicing before feeding the features into a window-level Q-Former. WavLLM [7] combines Whisper and WavLM [12] for feature extraction, with feature fusion accomplished through 1D convolution, bottleneck adapters, and linear projection structures. BLSP-Emo [13] use convolution-based subsampler as the modality adapter. In addition, AffectGPT as a multimodal model [14] uses a pre-fusion projector to fuse audio and video features.

However, these methods still have limitations in the fusion approach, mainly in the form of inadequate feature interaction. Most ALLMs use simple splicing or linear projection for fusion, and the semantic and acoustic information are loosely linked, making it difficult to capture complex dependencies. Moreover, splicing or linear transformation may lead to information loss and affect the alignment effect when dealing with features of different time scales. Furthermore, these ALLMs do not consider the contextual information, which makes them hard to distinguish emotions [8, 7]. Therefore, we are motivated to introduce EAA, an advanced approach for enhancing the emotion perception capabilities of ALLMs by incorporating dual cross-attention and context-aware instruction tuning.

3. EAA Design

The architecture of EAA is illustrated in Figure 1. The input audio is first processed by two distinct encoders: a semantic encoder and an acoustic encoder, each capturing complementary representations of the speech signal. These representations are then integrated using a dual cross-attention fusion module, which effectively aligns and enhances the interaction between semantic and acoustic features. To further refine emotion recognition, we employ context-aware instruction tuning, fine-tuning the LLaMA model with LoRA [15] for efficient adaptation. This approach enables the model to consider both the current and immediately preceding utterances, leveraging contextual information to enhance emotion recognition. Finally, the audio language model generates predicted emotion using the learnt representations and contextual information.

3.1. Dual Cross-attention Fusion for Audio Representation

To effectively integrate semantic and acoustic features, we design a dual cross-attention fusion module that enhances the interaction between these two representations. Given an input audio signal x, we employ two distinct encoders: the semantic encoder f_s and the acoustic encoder f_a.

The semantic encoder f_s (HuBERT [16]) leverages pseudo-labels generated from spectral features or the model's own learned representations, which have been shown to contain rich semantic information [17]. Meanwhile, the acoustic encoder f_a (BEATs [18]) focuses on capturing fine-grained spectral and prosodic characteristics directly from the raw audio signal.

The outputs of these encoders are represented as:

$$S = f_s(x) \in \mathbb{R}^{T_s \times d_s}, \quad A = f_a(x) \in \mathbb{R}^{T_a \times d_a} \tag{1}$$

where S and A represent the speech and acoustic features, T_s, T_a are sequence lengths, and d_s, d_a are the respective feature dimensions. To align these features in a shared latent space, we apply linear projections:

$$\tilde{S} = W_s S + b_s, \quad \tilde{A} = W_a A + b_a \tag{2}$$

where W_s, W_a are trainable projection matrices that map the features to a common dimension d. Before performing cross-attention, we apply layer normalization to both projected features $\tilde{S}$ and $\tilde{A}$ to stabilize training. Since the sequence lengths

T_s and T_a may differ, we pad the shorter sequence along the temporal axis with zeros, aligning both sequences to the same length $T = \max(T_s, T_a)$. This ensures that the attention mechanisms operate over temporally aligned representations.

Then, we employ dual cross-attention where each representation attends to the other. Specifically, the semantic-to-acoustic attention allows the semantic features to query the acoustic representations, incorporating detailed acoustic characteristics, while the acoustic-to-semantic attention enables the acoustic features to query the semantic representations, enriching them with linguistic information. Formally, given semantic features $\tilde{S}$ and acoustic features $\tilde{A}$, the attention mechanisms are defined as:

$$\begin{aligned} \text{Att}_s &= \text{Softmax}\left(\frac{Q_s K_a^T}{\sqrt{d}}\right) V_a \\ \text{Att}_a &= \text{Softmax}\left(\frac{Q_a K_s^T}{\sqrt{d}}\right) V_s \end{aligned} \tag{3}$$

where Q, K, and V are query, key, and value projections, respectively, and d represents the feature dimension. We use a simple yet effective concatenation strategy to fuse the original and attended features into the final representation:

$$F = \text{Concat}(\tilde{S}, \tilde{A}, \text{Att}_s, \text{Att}_a) \tag{4}$$

The fused representation F is then projected into the hidden space of the language model for downstream processing. This bidirectional interaction enhances the fusion of semantic and acoustic information, leading to a more comprehensive audio representation.

3.2. Context-aware Instruction Tuning

Identifying emotions from isolated speech is inherently challenging due to the absence of contextual cues. In natural dialogue, emotions evolve continuously, with prior discourse shaping the emotional state of the current utterance. To overcome this limitation, we propose a context-aware instruction tuning mechanism that effectively integrates contextual information into the emotion recognition process, enabling a more accurate and nuanced understanding of speech emotions.

Given an input utterance of a speech, we retrieve its most recent preceding utterance within the same conversation to form a contextual representation. Here, an utterance refers to a single complete sentence within the dialogue. Specifically, for an audio sample x_t at time step t, if a prior utterance x_{t-1} exists within the same dialogue, we concatenate them using a separator token:

$$C_t = x_{t-1} \oplus x_t \tag{5}$$

where $\oplus$ denotes sentence-level concatenation, allowing the model to capture contextual dependencies. If no previous utterance is available, the current utterance is used as a standalone input.

To further enhance the model's ability to recognize emotions, we employ instruction tuning with the prompt: *"Describe the speaker's emotion in one word."* This instruction ensures that the model's output aligns with the expected emotion categories. The final input to LLaMA consists of fused audio features, the instruction prompt, and contextual information, providing a richer representation for emotion recognition.

4. Performance Evaluation

The audio signals are processed at a sampling rate of 16kHz to maintain consistency across all samples. Each waveform is converted to a single-channel format, with resampling applied when necessary. Feature extraction is performed using two encoders: the semantic encoder HuBERT and the acoustic encoder BEATs. To balance efficiency and performance, all layers of both encoders are frozen except for the last two, which are fine-tuned to adapt to the emotion recognition task.

For efficient fine-tuning of the LLaMA-3-8B model, we employ LoRA, which updates only a subset of parameters while keeping the majority of the pre-trained model frozen. Specifically, we set the LoRA rank to 2, the LoRA scaling factor (α) to 16, and apply a dropout rate of 0.2 to enhance generalization. The model is trained with a batch size of 2 and a hidden size of 768. The optimizer used is AdamW with an initial learning rate of 5×10^{-6}. A linear warm-up is applied at the beginning of training, followed by a cosine decay learning rate schedule to stabilize optimization. The model is trained on a single H100 GPU.

Table 1: *Comparison of different audio-language models on emotion recognition. Any results that are directly cited from the original paper are denoted with the symbol †. Results reported by [20] are denoted with the symbol *, while those reported by [6] are marked with the symbol ‡. Results without any superscript were obtained from our own experiments.*

Model	Accuracy
Pengi [5]	0.289
MERaLiON † [10]	0.302
SALMONN * [8]	0.331
Whisper + Llama3 * [20]	0.334
WavLLM * [7]	0.411
WavLM-large ‡ [12]	0.542
Qwen2-audio † [21]	0.553
AffectGPT †[14]	0.557
Qwen-audio † [6]	0.557
OSUM † [11]	0.566
BLSP-Emo † [13]	0.573
EAA (Ours)	**0.687**

4.1. Dataset and Evaluation Metric

In this work, we utilize the MELD dataset [19] for training, validation, and testing because of its comprehensive annotations and widespread adoption in both pretraining and downstream emotion recognition tasks. MELD is a multimodal dataset incorporating text, video, and audio modalities. Since our focus is on audio-language models, we utilize only the text and audio data. The dataset covers seven emotion categories: neutral, joy, sadness, anger, fear, disgust, and surprise, comprising 13,847 utterances from 407 speakers, spanning a total of 12.2 hours of English speech. It is pre-divided into training, validation, and testing sets, ensuring a standardized evaluation process. For evaluation, we assess accuracy by comparing the predicted emotion labels generated by our proposed models with the ground truth annotations provided in the dataset, ensuring a reliable performance assessment.

4.2. Results

We compare our proposed model with various ALLMs, including Pengi [5], MERaLiON [10], SALMONN [8], WAVLLM [7], WavLM-large [12], QWen2-audio [21], AffectGPT [14], Qwen-audio [6], OSUM [11], and BLSP-emo [11]. Among

Table 2: *Comparison of traditional emotion recognition methods. The accuracy values for all baseline models are reported from their respective original papers. "✓" indicates that the model utilizes the corresponding modality.*

Model	Text	Audio	Video	Accuracy
UniMSE [29]	✓	✓	✓	0.651
MPT-HCL [30]	✓	✓	✓	0.659
SDT [27]	✓	✓	✓	0.676
CFN-ESA [26]	✓	✓	✓	0.679
M2FNet [25]	✓	✓	✓	0.679
SACL-LSTM [28]	✓			0.679
Mamba-like Model [23]	✓	✓	✓	0.680
GS-MCC [22]	✓	✓	✓	0.681
DF-ERC [24]	✓	✓	✓	0.683
ELR-GNN [31]	✓	✓	✓	0.687
EAA (Ours)	✓	✓		**0.687**

these, AffectGPT [14] is a multimodal LLM. The results of this comparison are presented in Table 1 show that EAA achieves state-of-the-art performance in ALLMs for speech emotion recognition, with an accuracy of 68.7%. Notably, even when compared to ALLMs specifically designed for emotional support [13], our method outperforms them by a significant margin of 11.4%, proving its superior effectiveness.

Additionally, we evaluate our approach against traditional classification methods that do not use the audio encoder and large language model architecture, including GS-MCC [22], Mamba-like Model [23], DF-ERC [24], M2FNet [25], CFN-ESA [26], SDT [27], and SACL-LSTM [28]. The results for these methods are shown in Table 2, which demonstrates that EAA remains competitive with traditional speech emotion classification models. Notably, our method achieves comparable performance to the ELR-GNN model, which incorporates an additional video modality. As can be seen from Table 1 and Table 2, most ALLMs are not as effective when compared to traditional classification methods. This is likely due to the fact that classification requires the model to choose from only seven emotion categories, whereas generation demands selecting an appropriate emotion-related word from a much larger vocabulary. Overall, our method enables ALLMs to achieve accuracy comparable to traditional classification models, effectively closing the performance gap.

Table 3: *Impact of contextual information on emotion recognition.*

Context Setting	Accuracy
No utterance (audio only)	0.523
Only current utterance	0.667
Current utterance + preceding sentence	0.687

4.3. Ablation Study

To further evaluate the effectiveness of our proposed contributions, we conduct a series of experiments. First, we investigated the effect of contextual information on emotion recognition ability in EAA. The results, presented in Table 3, highlight the impact of incorporating context on model performance. In this table, "No utterance" refers to the setting where neither the current utterance of the audio nor the preceding sentence from the previous audio is provided, i.e, there is no additional text information about the input audio. The results show that including only the current utterance significantly improves the model's ability to recognize emotions. Furthermore, adding the preceding sentence from the previous audio further enhances performance, emphasizing the importance of contextual information in audio-based emotion recognition.

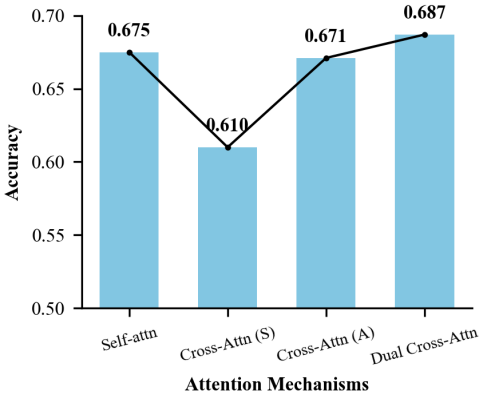

Figure 2: *Comparison of attention mechanisms.*

Next, we evaluate the effectiveness of the dual cross-attention mechanism by comparing it with alternative attention strategies. Specifically, we consider three additional configurations: (1) a model that relies solely on self-attention, where acoustic and semantic features independently undergo self-attention without any interaction between them; (2) a single cross-attention approach where acoustic features serve as the query while semantic features act as the key and value; and (3) a reverse setup where semantic features serve as the query while acoustic features function as the key and value. In all configurations, both the current utterance and the preceding sentence are included to provide contextual information. The results are presented in Figure 2. As observed, the model performs better when acoustic features are used as queries and semantic features serve as keys and values, compared to the reverse setting where semantic features are queries. This indicates that cross-attention with acoustic features attending to original semantic representations captures more useful information for emotion recognition than the opposite direction. Using self-attention alone can still achieve relatively good performance, but it is not as effective as dual cross-attention in feature fusion.

5. Conclusions

This study investigates the emotion-awareness capability of audio large language models. We propose EAA (Emotion-Aware Audio LLMs), which integrates dual cross-attention fusion and context-aware instruction tuning to enhance emotion recognition. First, we employ a dual cross-attention mechanism to effectively fuse acoustic and semantic features while preserving the original un-fused features to ensure a more comprehensive audio representation. Then, recognizing the challenge of identifying emotions from isolated audio, we incorporate context-aware instruction tuning to fine-tune LLaMA, leveraging conversational context to improve the understanding of emotion. Experimental results demonstrate that EAA achieves superior performance in generating accurate emotion labels. In the future, we will explore generating supportive responses and improving generalization across emotions and contexts.

6. References

[1] C. Singla, S. Singh, P. Sharma, N. Mittal, and F. Gared, "Emotion recognition for human–computer interaction using high-level descriptors," *Scientific Reports*, vol. 14, p. 12122, 2024.

[2] N. Elsayed, Z. ElSayed, N. Asadizanjani, M. Ozer, A. Abdelgawad, and M. Bayoumi, "Speech emotion recognition using supervised deep recurrent system for mental health monitoring," in *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*. IEEE, 2022, pp. 1–6.

[3] Y. Feng and L. Devillers, "End-to-end continuous speech emotion recognition in real-life customer service call center conversations," in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2023, pp. 1–8.

[4] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023, accessed: 2025-02-15. [Online]. Available: https://arxiv.org/abs/2303.08774

[5] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 090–18 108, 2023.

[6] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[7] S. Hu, L. Zhou, S. Liu, S. Chen, L. Meng, H. Hao, J. Pan, X. Liu, J. Li, S. Sivasankaran *et al.*, "Wavllm: Towards robust and adaptive speech large language model," *arXiv preprint arXiv:2404.00656*, 2024.

[8] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," in *International Conference on Learning Representations (ICLR)*, 2024.

[9] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] Y. He, Z. Liu, S. Sun, B. Wang, W. Zhang, X. Zou, N. F. Chen, and A. T. Aw, "Meralion-audiollm: Technical report," *arXiv preprint arXiv:2412.09818*, 2024.

[11] X. Geng, K. Wei, Q. Shao, S. Liu, Z. Lin, Z. Zhao, G. Li, W. Tian, P. Chen, Y. Li, P. Guo, M. Shao, S. Wang, Y. Cao, C. Wang, T. Xu, Y. Dai, X. Zhu, Y. Li, L. Zhang, and L. Xie, "Osum: Advancing open speech understanding models with limited resources in academia," *arXiv preprint arXiv:2501.13306*, 2025.

[12] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[13] C. Wang, M. Liao, Z. Huang, J. Wu, C. Zong, and J. Zhang, "Blsp-emo: Towards empathetic large speech-language models," *arXiv preprint arXiv:2406.03872*, 2024.

[14] Z. Lian, H. Chen, L. Chen, H. Sun, L. Sun, Y. Ren, Z. Cheng, B. Liu, R. Liu, X. Peng *et al.*, "Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models," *arXiv preprint arXiv:2501.16566*, 2025.

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021, arXiv preprint arXiv:2106.09685. [Online]. Available: https://arxiv.org/abs/2106.09685

[16] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[17] T. Maekaku, J. Shi, X. Chang, Y. Fujita, and S. Watanabe, "Hubertopic: Enhancing semantic representation of hubert through self-supervision utilizing topic model," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 741–11 745.

[18] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023, pp. 5178–5193. [Online]. Available: https://proceedings.mlr.press/v202/chen23ag.html

[19] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 527–536. [Online]. Available: https://aclanthology.org/P19-1050/

[20] B. Wang, X. Zou, G. Lin, S. Sun, Z. Liu, W. Zhang, Z. Liu, A. Aw, and N. F. Chen, "Audiobench: A universal benchmark for audio large language models," *arXiv preprint arXiv:2406.16020*, 2024.

[21] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Lu, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[22] T. Meng, F. Zhang, Y. Shou, W. Ai, N. Yin, and K. Li, "Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum," *arXiv preprint arXiv:2404.17862*, 2024.

[23] Y. Shou, T. Meng, F. Zhang, N. Yin, and K. Li, "Revisiting multi-modal emotion learning with broad state space models and probability-guidance fusion," *arXiv preprint arXiv:2404.17858*, 2024.

[24] B. Li, H. Fei, L. Liao, Y. Zhao, C. Teng, T.-S. Chua, D. Ji, and F. Li, "Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5923–5934.

[25] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2fnet: Multi-modal fusion network for emotion recognition in conversation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4652–4661.

[26] J. Li, X. Wang, Y. Liu, and Z. Zeng, "Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition," *IEEE Transactions on Affective Computing*, 2024.

[27] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, "A transformer-based model with self-distillation for multimodal emotion recognition in conversations," *IEEE Transactions on Multimedia*, vol. 26, pp. 776–788, 2024.

[28] D. Hu, Y. Bao, L. Wei, W. Zhou, and S. Hu, "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)*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 10 835–10 852. [Online]. Available: https://aclanthology.org/2023.acl-long.606/

[29] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022, pp. 787–801. [Online]. Available: https://aclanthology.org/2022.emnlp-main.534.pdf

[30] S. Zou, X. Huang, and X. Shen, "Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 5994–6003.

[31] Y. Shou, W. Ai, J. Du, T. Meng, H. Liu, and N. Yin, "Efficient long-distance latent relation-aware graph neural network for multi-modal emotion recognition in conversations," *arXiv preprint arXiv:2407.00119*, 2024.