

EII-SCL: Harnessing Emotional Inertia for Multimodal Emotion Recognition in Conversation

Zilong Huang[✉], Kong Aik Lee[✉], Chong-Xin Gan, Zezhong Jin, Ruichen Zuo, Man-Wai Mak[✉]

Dept. of Electrical and Electronic Engineering,
The Hong Kong Polytechnic University, Hong Kong SAR, China
{zi-long.huang, kong-aik.lee, man.wai.mak}@connect.polyu.hk

Abstract

Multimodal emotion recognition in conversation (MERC) achieves accurate predictions by integrating multimodal and contextual information in dialogues. While current MERC approaches focus on modeling complex contextual dependencies in conversation, they often overlook the impact of contextual emotional inertia in emotion shift, leading to sub-optimal performance. To address this issue, we propose a novel **Emotional Inertia-Informed Supervised Contrastive Learning** module (EII-SCL) that informs the contrastive objective by constructing inertia-affected samples within temporal windows, effectively leveraging emotional inertia as a prior while enabling seamless integration with existing MERC models without requiring additional data. Extensive experiments on IEMOCAP and MELD show that our approach consistently outperforms state-of-the-art methods.

Index Terms: Emotional Inertia, Emotion Recognition in Conversation, Multimodal Network, Contrastive Learning

1. Introduction

Emotion recognition in conversation (ERC) is a key technique in applications such as human-computer interaction and intelligent medical care [1]. Different from traditional emotion recognition, ERC analyzes dialogue turns and assigns one emotion label per utterance in the dialogue [2, 3].

The core of ERC lies in modeling contextual dependencies between utterances. Mainstream methods construct dialogue structures through graph neural networks to explicitly capture such a relationship. MMGCN [4] extends graph neural networks to multimodal settings, exploring a more effective way of utilizing both multimodal and long-distance contextual information. AdaIGN [5] uses an adaptive graph selection strategy to balance intra- and inter-modal interaction, better handling the challenge of balancing cross-speaker contextual dependencies. Recent studies have explored advanced representation learning strategies to improve feature discrimination in conversational and speech-related tasks [6–10]. However, most of the contextual modeling works [11–14] overlook emotional inertia, making their models unable to leverage such an important psychological behavior of humans to improve ERC performance.

Emotional inertia [15, 16] is a psychological concept that describes the resistance of emotions to change. In a conversational environment, an individual’s previous emotional state will have a lasting impact on current and future emotions, i.e., the emotional state of a speaker tends to maintain a gradual and smooth transition rather than irregular and drastic changes. Current research generally studies the phenomenon of emotion shift between discourses, such as introducing emotion shift detection modules [17], utilizing conversation-level curriculum learning

based on emotion shift frequency to progressively enhance the model’s ability to recognize emotions [18], and modeling multiple emotion labels [19, 20]. Although these studies have successfully modeled the emotion shift phenomenon, they failed to fully consider the emotional dependence of the same speaker in continuous emotion expressions and emotion transitions, which limits the in-depth understanding of the dynamics of emotional inertia.

To address the above challenges, we propose a module called *emotional inertia-informed supervised contrastive learning* (EII-SCL). Specifically, we introduce a supervised contrastive loss, which imposes constraints on multimodal features based on the emotional inertia of the same speaker within a local time window. As a general module, EII-SCL can be integrated into typical MERC models and improve their performance without adding additional data annotations. Compared with existing work, our main contributions are reflected in three aspects:

- This paper proposes an innovative supervised contrastive learning module called EII-SCL. The module constructs hard-negative samples influenced by emotional inertia by jointly leveraging speaker identity, emotional labels, and local time windows, and it designs a dynamic weighting mechanism for these samples to capture the gradual transitions between different emotional states.
- By rationally selecting positive, easy-negative and hard-negative samples, we analyze and validate the emotional inertia phenomenon on emotion recognition datasets.
- Our experiments demonstrate that EII-SCL can harness emotional inertia for MERC tasks without adding additional data. A basic MERC model equipped with the EII-SCL module can outperform the current SOTA methods.

2. Methodology

2.1. Task Definition

Consider a dialogue consisting of k temporally ordered utterances, denoted as $\mathcal{U} = \{u_i\}_{i=1}^k$, where individual utterances have corresponding emotion labels $\{y_i\}_{i=1}^k$ and speaker labels $\{s_i\}_{i=1}^k$. The emotion label y_i belongs to a predefined set of emotions $\mathcal{Y}$. Each utterance u_i is accompanied by multimodal data, including a video clip, an audio segment, and a text transcript. Formally, we represent a dialogue as follows:

$$\mathcal{U} \times \mathcal{Y} \times \mathcal{S} = \left\{ \{u_i^{(\delta)}, y_i, s_i\} \mid \delta \in \{a, t, v\}, y_i \in \mathcal{Y}, s_i \in \mathcal{S} \right\}, \quad (1)$$

where $u_i^{(\delta)}$ denotes the modality-specific representation (text, audio, or video) of the i -th utterance. The ERC task involves predicting the emotion label $\hat{y}_i$ for each utterance u_i in the dialogue, leveraging information from all k utterances.

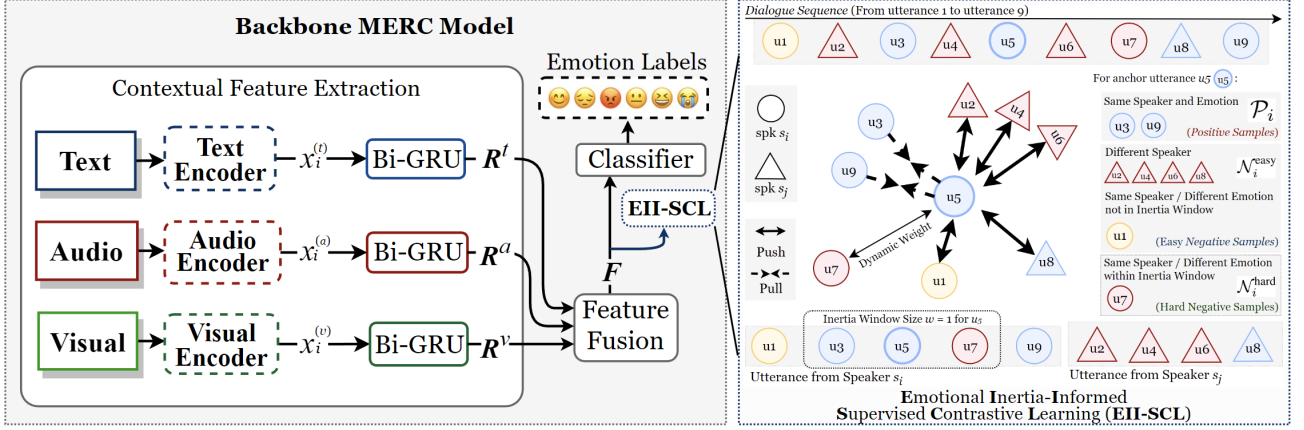


Figure 1: The overall architecture of the proposed EII-SCL approach. The backbone MERC model completes emotion prediction through contextual features extraction, multimodal feature fusion, and emotion classification. EII-SCL works with the backbone network to learn more subtle and discriminative feature representations through a carefully designed inertia-informed sampling strategy. The thickness of the lines indicates the intensity of the repulsive force, and different colors of the utterances represent different emotion labels. Solid-line and dashed-line rectangles represent trainable and frozen models, respectively.

2.2. Overall Framework

As shown in Figure 1, the proposed *emotional inertia-informed supervised contrastive learning* (EII-SCL) framework consists of the following components. First, a basic MERC model processes the conversation data and generates the final emotion predictions. The EII-SCL module processes the fused multimodal features. Through a supervised contrast loss function, the module improves the perception of emotional inertia of the MERC and feature fusion models through supervised negative samples construction.

2.3. Backbone MERC Model

A basic MERC model typically consists of three parts: contextual feature extraction, multimodal feature fusion, and emotion classification. For each modality, we employ pre-trained encoders to obtain utterance-level representations:

$$\mathbf{x}_i^{(\delta)} = \mathbf{f}_{\text{enc}}^{(\delta)}(u_i^{(\delta)}), \delta \in \{t, a, v\}, i = 1, \dots, k, \quad (2)$$

where $\delta \in \{t, a, v\}$ and the modality-specific embedding is obtained by averaging the final layer’s outputs of the corresponding encoder. To capture the temporal dependencies within a dialogue, the modality-specific embeddings are processed using bidirectional GRUs:

$$\mathbf{R}^{(\delta)} = [\mathbf{r}_1^{(\delta)}, \dots, \mathbf{r}_k^{(\delta)}] = \text{Bi-GRU}(\mathbf{x}_1^{(\delta)}, \dots, \mathbf{x}_k^{(\delta)}). \quad (3)$$

The contextual representations from different modalities are fused via a feature fusion module (e.g., Transformer-based [21] or Graph-based [22]):

$$\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_k] = \text{FeatureFusion}[\mathbf{R}^{(t)}, \mathbf{R}^{(a)}, \mathbf{R}^{(v)}], \quad (4)$$

where $\mathbf{f}_i \in \mathbb{R}^{D_f}$ denotes the fused multimodal embedding of utterance u_i . Each fused embedding $\mathbf{f}_i$ is fed into a classifier to obtain emotion probabilities $\mathbf{p}_i$, optimized using the standard cross-entropy loss $\mathcal{L}_{CE}$.

2.4. Emotional Inertia-Informed Contrastive Learning

The key to modeling emotional inertia lies in effectively distinguishing between utterances influenced by emotional inertia

and those that are not. In this work, we apply the concept of emotional inertia in the context of contrastive learning. Previous supervised contrastive learning methods do not take into account the emotion labels of adjacent utterances, i.e., they only treat utterances with the same emotion label as positive pairs and those with different emotion labels as negative pairs, regardless of their locality in the dialog.

Emotional inertia reflects the temporal persistence and resistance to change of emotional states. This characteristic suggests an important observation: in a dialogue, even when an emotion shift occurs, adjacent utterances from the same speaker often remain semantically similar due to the gradual transition between emotional states, although they are assigned different discrete emotion labels. If a contrastive learning framework ignores such emotion transitions when forming negative pairs, it fails to account for the dynamic influence of emotional inertia, thereby limiting the model’s ability to capture the gradual evolution of emotional states.

To address the aforementioned issue, we redefine the positive and negative sets by incorporating an adaptive inertia window that dynamically adjusts the context scope based on the semantic similarity between utterances. The positive pairs $\mathcal{P}_i$ for the i -th utterance consist of samples from the same speaker with the same emotion label:

$$\mathcal{P}_i = \{\mathbf{f}_j \mid s_j = s_i, y_j = y_i, \forall j \neq i\}, \quad (5)$$

where s_i and y_i denote the speaker and emotion labels of utterance u_i , respectively. To adaptively determine the contextual scope of emotional inertia, we estimate the inertia window using an attention mechanism. Specifically, the attention weight between the anchor utterance i and another utterance j from the same speaker is computed as:

$$a_{i,j} = \frac{\exp((\mathbf{W}_q \mathbf{f}_i)^\top (\mathbf{W}_k \mathbf{f}_j))}{\sum_{\substack{t \neq i \\ s_t = s_i}} \exp((\mathbf{W}_q \mathbf{f}_i)^\top (\mathbf{W}_k \mathbf{f}_t))}, s_j = s_i \quad (6)$$

where $\mathbf{W}_q$ and $\mathbf{W}_k$ are learnable query and key projection matrices, and t indexes all utterances from the same speaker except the anchor i . The inertia window size is then defined as the

attention-weighted temporal distance:

$$\omega_i = \left\lfloor \sum_{j \neq i, s_j = s_i} a_{i,j} |i - j| \right\rfloor, \quad (7)$$

where $|i - j|$ denotes the index-based temporal distance between the j^{th} utterance and i^{th} utterance spoken by the same speaker, and $\lfloor \cdot \rfloor$ converts the result to an integer window size. For the negative set, we define the emotional inertia window $\mathcal{W}_i$ as:

$$\mathcal{W}_i = \{j \mid s_j = s_i, \forall |i - j| \leq \omega_i\}, \quad (8)$$

By using the emotional inertia window, we construct two distinct negative sample sets: $\mathcal{N}_i^{\text{easy}}$ and $\mathcal{N}_i^{\text{hard}}$:

$$\begin{aligned} \mathcal{N}_i^{\text{easy}} &= \{f_j \mid s_j \neq s_i\} \cup \{f_j \mid s_j = s_i, y_j \neq y_i, j \notin \mathcal{W}_i\} \\ \mathcal{N}_i^{\text{hard}} &= \{f_j \mid s_j = s_i, y_j \neq y_i, j \in \mathcal{W}_i\}. \end{aligned} \quad (9)$$

The easy-negative set $\mathcal{N}_i^{\text{easy}}$ combines utterances from different speakers with those from the same speaker but different emotion labels and falls outside the inertia window $\mathcal{W}_i$. The hard-negative set $\mathcal{N}_i^{\text{hard}}$ captures the emotional states of same-speaker utterances with different emotion labels within the inertia window $\mathcal{W}_i$, reflecting the emotional persistence characteristic of emotional inertia. This formulation models emotional inertia by considering speaker-specific temporal windows.

Given the fused embeddings $\{f_i\}_{i=1}^k$, we define the emotional inertia-informed contrastive loss as:

$$\begin{aligned} S_i^+ &= \sum_{f_j \in \mathcal{P}_i} \exp(\cos(f_i, f_j)/\tau), \\ S_i^- &= \sum_{f_t \in \mathcal{N}_i^{\text{easy}}} \exp(\cos(f_i, f_t)/\tau) \\ &+ \sum_{f_h \in \mathcal{N}_i^{\text{hard}}} \underbrace{\left(\frac{1 - \cos(f_i, f_h)}{2} \right)}_{\text{dynamic weight}} \cdot \exp(\cos(f_i, f_h)/\tau), \end{aligned} \quad (10)$$

where $\cos(\cdot)$ denotes the cosine similarity and τ is a temperature parameter. The dynamic weight adjusts the repulsive force based on their similarity to the anchor, thereby avoiding over-penalizing potentially similar utterances while maintaining discriminability. The final contrastive loss for emotional inertia modeling is computed as:

$$\mathcal{L}_{eii} = -\frac{1}{k} \sum_{i=1}^k \log \frac{S_i^+}{S_i^+ + S_i^-}. \quad (11)$$

2.5. Joint Training

The proposed model is optimized with a loss function $\mathcal{L}$ that combines the basic cross-entropy loss $\mathcal{L}_{CE}$ with $\mathcal{L}_{eii}$ for emotional inertia-informed learning. Specifically, EII-SCL collaborates with the Basic MERC model by jointly minimizing the following losses:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{eii}, \quad (12)$$

where α is the coefficient used to balance the two losses.

3. Experiments

3.1. Datasets and Evaluation Metric

IEMOCAP [25] is a widely used dataset for emotion recognition in conversation. For partitioning the data, we utilized the commonly used LOSO (Leave-One-Session-Out) strategy. **MELD** [26] is a multi-modal, multi-speaker conversational dataset derived from the TV series Friends. To ensure a fair comparison, we followed the predefined train/val/test splits provided by the dataset, ensuring data allocation aligned with [27]. Similar to other recent studies, we adopt the accuracy and weighted F1-score (w-F1) for evaluation.

3.2. Backbone MERC Models and Implementation Details

In order to measure the effectiveness of our proposed EII-SCL module, we constructed our backbone model through two representative MERC methods: Transformer-based and Graph-based methods.

The Transformer-based method uses the concatenated features through a transformer encoder to further capture multi-modal complementary information, which we refer to as MM-TransFormer:

$$\mathbf{F} = [f_1, \dots, f_k] = \text{TransEncoder}[\mathbf{R}^t, \mathbf{R}^a, \mathbf{R}^v], \quad (13)$$

The Graph-based method refers to the graph construction approach of DialogueGCN [22]. Given that this model primarily focuses on the single-text modality, we extend it to a multi-modal framework for a fair comparison. Specifically, we concatenate the multimodal features as the input, construct the graph, and perform feature transformation in a manner consistent with the original configurations, which we refer to as MM-DialogueGCN:

$$\mathbf{F} = [f_1, \dots, f_k] = \text{DialogueGCN}[\mathbf{R}^t, \mathbf{R}^a, \mathbf{R}^v], \quad (14)$$

We integrate the proposed EII-SCL into these two backbone MERC models to evaluate the impact of EII-SCL.

To ensure that methods remain consistent at the feature extraction level, we used the same pre-trained encoder setting to extract features of the three modalities in the backbone MERC models. Specifically, we use RoBERTa [28], Wav2vec2.0 [29], and CLIP [30] for the text, audio, and visual modalities, respectively. We conducted all experiments on a single NVIDIA RTX 4090 GPU, training for 40 epochs on IEMOCAP and 50 epochs on MELD with a batch size of 13. The model was optimized using Adam with a learning rate of 0.0002 and a dropout rate of 0.3. The temperature $\tau = 0.07$ and scaling factors $\alpha = 0.02$. Finally, we averaged the weights from the last 10 checkpoints to obtain the final model for evaluation.

3.3. Overall Results

Table 1 shows the detailed results on both datasets. After integrating the EII-SCL module, both backbone networks significantly outperform all baseline models. Notably, while the FEMI [12] shows outstanding performance in recognizing “frustrated” and “sad” emotions, its w-F1 score for “happy” is significantly lower (60.60%), revealing potential emotional label bias in conventional approaches.

3.4. Emotional Inertia Analysis in IEMOCAP Dataset

To explore how emotional inertia affects emotion recognition in conversation, we analyze this phenomenon in the IEMOCAP dataset. We first define hard-negative set ($\mathcal{N}_i^{\text{hard}}$) and

Table 1: Accuracy and weighted-average F1 score (w-F1) compared with other baselines. Best and second-best results are displayed in **bold** and underline, respectively. “MM-TransFormer” means using a simple Transformer encoder as the feature fusion module to fuse the concatenated multimodal features. * indicates our reproduced results from DialogGCN under the multimodal setting (T+A+V).

Baseline Model	Proposed Year	IEMOCAP								MELD	
		Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc	w-F1	Acc	w-F1
DialogGCN [22]	2019	42.75	84.54	63.54	64.19	63.08	66.99	65.25	64.18	-	58.10
MMGCN [4]	2021	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26	60.42	58.31
CFN-ESA [23]	2023	53.67	80.60	<u>71.65</u>	70.32	74.82	68.06	71.04	70.78	<u>67.85</u>	66.70
AdaIGN [5]	2024	53.04	81.47	71.26	65.87	76.34	67.79	-	70.74	-	66.79
DER-GCN [24]	2025	58.80	79.80	61.50	<u>72.10</u>	73.30	67.80	69.70	69.40	66.80	66.10
FEMI [12]	2025	60.60	85.55	70.92	70.98	75.13	69.00	71.97	<u>73.53</u>	64.88	66.41
MM-DialogGCN*		73.53	74.14	67.35	71.12	73.48	71.53	71.45	71.14	67.32	66.28
-with EII-SCL (Proposed)	2026	76.92	65.83	69.42	72.24	<u>78.42</u>	73.26	<u>73.13</u>	73.15	67.83	<u>66.97</u>
MM-TransFormer		73.84	67.48	70.90	71.18	75.56	72.43	72.53	72.57	67.64	66.58
-with EII-SCL (Proposed)	2026	<u>75.81</u>	77.59	72.16	69.10	79.09	<u>72.79</u>	73.95	74.01	68.19	67.33

easy-negative set ($\mathcal{N}_i^{\text{easy}}$) using different window sizes ω . As illustrated in Figure 2, the proportion of hard-negative set among all negative samples increases rapidly with larger window sizes, indicating that more negative samples are identified as hard-negative samples. Simultaneously, we evaluate the aver-

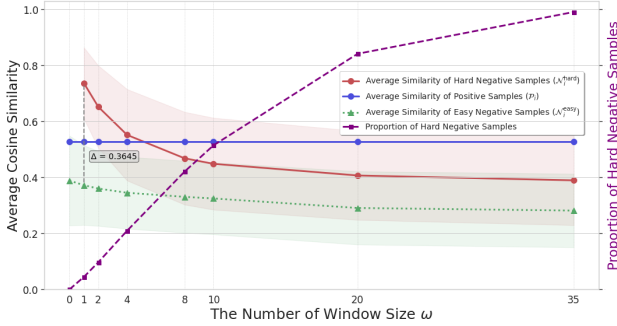


Figure 2: Impact of window size ω on the average similarity and proportion of negative samples on the IEMOCAP dataset. The proportion of hard-negative samples indicates the ratio of hard-negative samples to all negative samples. When $\omega = 0$, all negative samples are considered as easy-negative samples.

age cosine similarity between hard-negative and easy-negative pairs, respectively. The results demonstrate that across all window size configurations, hard-negative samples consistently exhibit significantly higher average similarity compared to easy-negative samples (all p-values ≤ 0.001). Notably, the similarity gap is most pronounced at the smallest window size ($\omega = 1$), where the average similarity of hard-negative samples surpasses that of easy-negative samples by a substantial margin of 0.3645, and even the positive samples. This validates our hypothesis, making these inertia-affected utterances more challenging to distinguish in the typical MERC methods.

3.5. Impact of Dynamic Inertia Window

The results in Figure 3 compare the proposed dynamic inertia window with fixed window settings. Fixed windows introduce a predefined temporal scope that may not accurately reflect the varying persistence of emotional states in conversations. In contrast, the proposed attention-based dynamic window size ω^i adaptively adjusts the contextual range according to the semantic relevance between utterances. This adaptive mechanism enables the model to capture emotional inertia more effectively by expanding the window in emotionally stable regions while contracting it during abrupt emotional transitions. As a result, the

dynamic window consistently provides more informative hard-negative samples, leading to improved ERC performance compared with fixed window configurations.

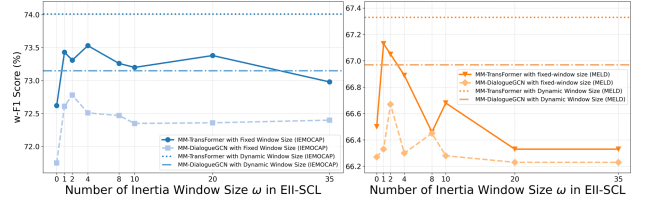


Figure 3: Impact of the dynamic and fixed inertia window size on ERC performance. The dynamic window adjusts to capture emotional persistence for each utterance.

3.6. Performance on Ambiguous Emotion Pairs

The results in Table 2 indicate that the proposed EII-SCL module effectively reduces misclassification rates between ambiguous emotion pairs (e.g., *Hap-Exc* means the true “Happy” label was incorrectly predicted as “Excited”), thereby improving the overall emotion prediction performance. This demonstrates EII-SCL’s capability to enhance emotion discriminability by learning more separable representations in confusing scenarios.

Table 2: Misclassification rates reduction of EII-SCL between ambiguous emotion pairs on IEMOCAP (Lower the better).

Methods	Exc-Hap	Sad-Fru	Neu-Fru	Ang-Fru	Fru-Neu
MM-DialogGCN	8.85	19.23	19.56	27.37	12.22
-with EII-SCL	6.36	19.88	14.08	24.54	11.92
MM-TransFormer	7.54	24.62	16.53	19.55	8.98
-with EII-SCL	5.37	19.05	14.78	17.35	8.11

4. Conclusions

We have proposed the emotional inertia-informed supervised contrastive learning (EII-SCL), which leverages emotional inertia—the psychological phenomenon in which emotional states tend to resist change over time—to enhance multimodal emotion recognition in conversation. Specifically, we select hard-negative samples within speaker-specific temporal windows in the supervised contrastive learning, enabling the model to learn more discriminative feature representations. EII-SCL can be seamlessly integrated into the backbone model to perceive the inherent emotional inertia in conversations, demonstrating state-of-the-art performance on both the MELD and IEMOCAP datasets.

5. Acknowledgment

The work presented in this article is supported by the Research Platform for Advanced Audio and Speech Signal Processing (P0049192) funded by Innovation Technology Co. Ltd.

6. Use of Generative AI Disclosure

Generative AI tools were used only for language polishing and formatting assistance. All scientific content, experiments and analyses were produced and verified by the authors.

7. References

- [1] A. Yadollahi, A. G. Shahraiki, and O. R. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–33, 2017.
- [2] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.
- [3] Z. Huang, M.-W. Mak, and K. A. Lee, "MM-NodeFormer: Node transformer multimodal fusion for emotion recognition in conversation," in *Proc. Interspeech*, 2024, pp. 4069–4073.
- [4] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," *arXiv preprint arXiv:2107.06779*, 2021.
- [5] G. Tu, T. Xie, B. Liang, H. Wang, and R. Xu, "Adaptive graph learning for multimodal conversational emotion detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/29876>
- [6] Z. Jin, S. Liu, Z. Li, C.-X. Gan, Z. Huang, M.-W. Mak, and K. A. Lee, "Distilling attention knowledge for speaker verification," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 16 447–16 451.
- [7] C.-X. Gan, P. Bell, M.-W. Mak, Z. Li, Z. Jin, Z. Huang, and K. A. Lee, "UNet-based fusion and exponential moving average adaptation for noise-robust speaker recognition," *arXiv preprint arXiv:2604.25624*, 2026.
- [8] C.-B. Nguyen, D.-T. Le, Q. T. Ha *et al.*, "Curriculum learning meets directed acyclic graph for multimodal emotion recognition," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 4259–4265.
- [9] Z. Jin, Y. Tu, Z. Li, Z. Huang, C.-X. Gan, and M.-W. Mak, "Denoising student features with diffusion models for knowledge distillation in speaker verification," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [10] W. Ren, K. Li, R. Chao, J. Li, Z. Huang, S. Ahmed, Y.-J. Li, K.-H. Hung, S.-S. Wang, H.-M. Wang *et al.*, "BAV-MossFormer2: Enhanced MossFormer2 for binaural audio-visual speech enhancement," in *Proc. AVSEC 2025*, 2025, pp. 79–80.
- [11] M. K. Tellamekala, S. Amiriparian, B. W. Schuller, E. André, T. Giesbrecht, and M. Valstar, "COLD Fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 805–822, 2024.
- [12] Y. Fu, X. Yan, W. Chen, and J. Zhang, "Feature-enhanced multimodal interaction model for emotion recognition in conversation," *Knowledge-Based Systems*, vol. 309, p. 112876, 2025.
- [13] C.-X. Gan, Z. Li, Z. Jin, Z. Huang, M.-W. Mak, and K. A. Lee, "IDIR: Identifying and Distilling Informative Relations for Speaker Verification," in *Interspeech 2025*, 2025, pp. 5758–5762.
- [14] 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, 2023.
- [15] P. Kuppens, N. Allen, and L. Sheeber, "Emotional inertia and psychological maladjustment," *Psychological science*, vol. 21, pp. 984–91, 07 2010.
- [16] P. Koval, A. Brose, M. Pe, M. Houben, Y. Erbas, D. Champagne, and P. Kuppens, "Emotional inertia and external events: The roles of exposure, reactivity, and recovery," *Emotion (Washington, D.C.)*, vol. 15, 04 2015.
- [17] Q. Gao, B. Cao, X. Guan, T. Gu, X. Bao, J. Wu, B. Liu, and J. Cao, "Emotion recognition in conversations with emotion shift detection based on multi-task learning," *Knowledge-Based Systems*, vol. 248, p. 108861, 04 2022.
- [18] L. Yang, Y. Shen, Y. Mao, and L. Cai, "Hybrid curriculum learning for emotion recognition in conversation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, 2022, pp. 11 595–11 603.
- [19] Joosung Lee, "The Emotion is Not One-hot Encoding: Learning with Grayscale Label for Emotion Recognition in Conversation," in *Proc. Interspeech*, 2022, pp. 141–145.
- [20] Y. Kang and Y.-S. Cho, "Beyond single emotion: Multi-label approach to conversational emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 24 321–24 329.
- [21] 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.
- [22] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," *arXiv preprint arXiv:1908.11540*, 2019.
- [23] J. Li, Y. Liu, X. Wang, and Z. Zeng, "CFN-ESA: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition," *arXiv preprint arXiv:2307.15432*, 2023.
- [24] W. Ai, Y. Shou, T. Meng, and K. Li, "DER-GCN: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 4908–4921, 2025.
- [25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.
- [26] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.
- [27] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7037–7041.
- [28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [29] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.
- [30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. International Conference on Machine Learning*, 2021, pp. 8748–8763.