From Discrete to Continuous: Progressive Hybrid-Distributional Learning for Gradual Emotion Transitions

Yunhe Xie and **Yang Li***

School of Computer Science and Artificial Intelligence, Northeast Forestry University

{xieyh, yli}@nefu.edu.cn

Abstract

Modeling emotional dynamics in multi-turn dialogues remains challenging due to emotion shifts, where emotional states evolve gradually rather than change abruptly. Existing methods often rely on one-hot supervision, which fails to reflect the progressive nature of human emotions, particularly for neutral-valence emotions with subtle trajectories. To address this limitation, we propose a pseudo soft-label guided Progressive Hybrid-Distributional (PHD) learning framework. PHD reconstructs discrete labels into hybrid distributional representations that encode inter-emotion relations and mixed emotional tendencies. Based on these representations, a progressive training strategy is introduced to guide the model from learning blended emotional states toward the standard one-to-one prediction objective. Furthermore, we design a Graded Contrastive Learning mechanism that replaces rigid binary allocation with graded supervision to alleviate label conflicts. Experiments on benchmark datasets demonstrate PHD consistently improves diverse baseline models, yielding average w-F1 gains of 1.29%. Our framework highlights the critical importance of modeling emotional dynamics as a gradual, distributional process.

1 Introduction

Endowing machines with complex emotional reasoning is a crucial step from artificial general intelligence toward social intelligence, driving extensive research attention on Emotion Recognition in Conversations (ERC) [Xu *et al.*, 2025]. ERC aims to track the emotional trajectory of interlocutors throughout an interaction. In real conversations, an interlocutor's emotion may fluctuate across dialogue turns, a phenomenon known as *emotion shift* [Zha *et al.*, 2025]. Current ERC solutions still exhibit a large performance gap between emotion shift scenarios and those involving stable-emotion, limiting overall reliability [Tu *et al.*, 2024; Yang *et al.*, 2025].

Prior work attributes this insensitivity to emotional shifts mainly to model's tendency to replicate frequent emotion labels from nearby utterances while overlooking the actual semantic cues [Xie *et al.*, 2025]. We argue that a more fundamental issue lies in the fact that **human emotional transitions are dominated by gradual evolution rather than abrupt jumps** [Kuppens *et al.*, 2010]. For example, in a conversation from the MELD [Poria *et al.*, 2019] dataset shown in Figure 1, Ross's `Anger` does not emerge as a sudden label-level jump, but progressively accumulates until it becomes dominant in utterance u_5. Conventional one-hot annotations can only capture superficial transition patterns and fail to reveal deeper gradual-evolution dynamics [Cambria *et al.*, 2026]. Hence, models can at best detect an emotion shift but struggle to infer where the shift is heading [Li *et al.*, 2025]. This limitation becomes particularly pronounced when inferring shifts involving neutral-valence emotions, such as `Surprise`, whose subtle trajectories cannot be adequately represented by discrete labels [Roohi *et al.*, 2025].

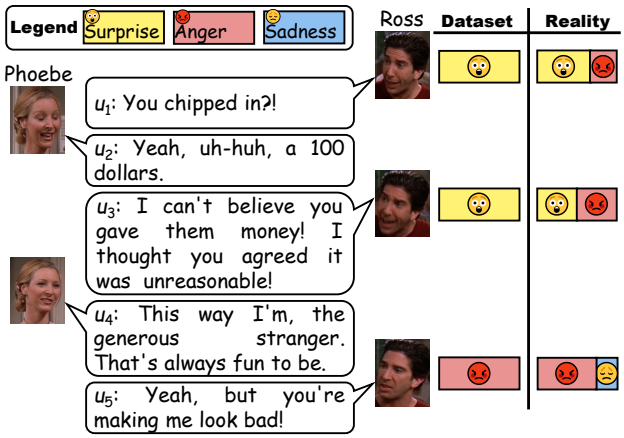

Figure 1: A dialogue excerpt from the MELD dataset. Ross's anger gradually intensifies throughout the conversation, while the one-hot annotations depict only an apparent label-level jump.

Moreover, allocating multiple emotion labels or even confidence scores per utterance during dataset construction heavily relies on annotators' subjective judgments and difficult to scale. Therefore, existing ERC models typically optimize a one-to-one matching objective that selects a most appropriate emotion label for each utterance [Pereira *et al.*, 2025]. This setting is inherently suboptimal as **human emotional**

*Corresponding author.

expressions are intrinsically entangled [Barrett, 2017], and focusing solely on the dominant emotion inevitably discards valuable complementary cues [Mao *et al.*, 2025].

To overcome the constraints of one-hot supervision, we propose a pseudo soft-label guided **P**rogressive **H**ybrid-**D**istributional (**PHD**) learning framework, which can be seamlessly integrated with existing ERC methods across different paradigms. Specifically, we first reconstruct one-hot emotion labels into *distributed representations* in both the static emotion space and the dynamic contextual space, and then hybridize them. The resulting soft labels encode both fundamental inter-emotion relations and the multifaceted emotional tendencies in the utterance. Building on these representations, we further introduce a novel *progressive training* strategy that gradually guides the model from capturing mixed emotional states to converging toward the one-to-one matching objective. Additionally, inspired by graded decision systems (e.g., multi-level paper scoring rules in certain academic conferences) rather than binary evaluation, we design a *Graded Contrastive Learning* (GCL) mechanism. Beyond traditional positive and negative sets, GCL further defines weak-positive and weak-negative sets according to the emotion distribution, allowing a controllable gray zone of exchangeable samples between positive and negative sets and allocating loss weights accordingly in a graded manner.

Our main contributions are summarized as follows:

1. We revisit the importance of distributional representations for modeling emotional dynamics in conversations and propose a cost-free soft-label generation method, ensuring strong portability of our PHD framework.
2. We propose a novel progressive training strategy that enables the model to mimic gradual emotion transition patterns, improving recognition performance in emotion shift scenarios, especially for neutral-valence emotions.
3. We design a graded contrastive mechanism that resolves label-conflict issues inherent in binary allocation rules, correcting the interference from outliers within positive/negative sets in conventional setups.

2 Related Work

2.1 Conversational Emotion Recognition

The primary feature of ERC lies in the context-dependent reasoning. Early studies employ Recurrent Neural Networks (RNNs) to sequentially encode utterances while struggle to capture the complex multi-party interaction [Hu *et al.*, 2021; Liu *et al.*, 2022]. Subsequent research increasingly adopts graph neural networks (GNNs) to propagate information. Some approaches treat utterances as nodes and construct fully-connected graphs [Joshi *et al.*, 2022; Zhang and Li, 2023]; others rely on discourse parsing to mitigate the noise from redundant edges [Sun *et al.*, 2021; Zhao *et al.*, 2022]. DualGAT executes both modeling schemes in parallel and introduces a differential regularizer to maximize the extraction of complementary information [Zhang *et al.*, 2023].

Beyond structural modeling, incorporating auxiliary information has proven effective. FATRER employs a topic modeling strategy that integrates emotion-aware signals [Mao *et al.*, 2023]. MKFM treats large language models (LLMs) as knowledge providers and analyzes their complementary contributions [Tu *et al.*, 2023b]. ERC-DP incorporates dynamic personality theory, introducing a personality prompt [Wang *et al.*, 2024]. The recent trend of instruction-tuning LLMs for specific tasks has also shown promise. DialogueLLM [Zhang *et al.*, 2025] uses ERNIE Bot [Baidu, 2023] to generate video–text descriptions and integrates them into instruction design. CoE advances this further by incorporating dialogue scenarios, and designs specialized multi-task instructions to incrementally fuse critical cues [Shen *et al.*, 2025b].

Despite these impressive advances, existing approaches still struggle in emotion-shift scenarios. They tend to adopt conservative strategies that replicate frequent labels, failing to robustly infer the direction of emotional transitions. This significantly hinders their applicability in real-world settings.

2.2 Emotion Shift Modeling

Enhancing model awareness of emotion-shift has recently become a research focus. One mainstream solution formulates emotion-shift detection as an auxiliary task [Yang *et al.*, 2025]. In addition, HCL defines the frequency of emotion shifts within a dialogue as a difficulty score and performs dialogue-level curriculum learning in an easy-to-hard manner [Yang *et al.*, 2022]. CLED interpolates emotion centroids to generate augmented utterances, simulating emotional dynamics based on the emotion transition matrix [Kang and Cho, 2024]. PCGNet treats consecutive utterance dyads as positive or negative samples according to their emotion-shift patterns to perform contrastive learning [Tu *et al.*, 2024].

The inherent limitation of one-hot supervision restricts current research to a relatively coarse level of emotion-shift perception, without capturing its direction. This limitation poses consistent challenges in modeling shifts involving neutral-valence emotions, such as transitions from the neutral state.

2.3 Interactive Dynamics Contrastive Learning

Complex interaction dynamics emerge from the information propagation among interlocutors, utterances, and their interactions. To better capture these dynamics, recent studies construct contrastive objectives. Specifically, positive and negative samples are defined based on whether utterances originate from the same dialogue [Nie *et al.*, 2023], are produced by the same interlocutor [Shen *et al.*, 2025a], or require auxiliary knowledge [Tu *et al.*, 2023a], thereby enhancing model sensitivity to context, identity, and semantics, respectively. Supervised contrastive learning grounded in explicit emotion labels remains the dominant paradigm and has demonstrated strong compatibility with a wide range of architectures [Dai *et al.*, 2024; Cao *et al.*, 2025; Ji *et al.*, 2025].

Existing contrastive frameworks largely rely on a rigid binary decision rule, which is suboptimal for emotion modeling. This all-or-nothing allocation can lead to the loss of nuanced emotional information. For instance, when *joy* is treated as positive, *excitement*, *neutral*, and *sadness* are all indiscriminately grouped together to the negative set, despite their different degrees of affective opposition. Current methods overlook such gradual emotional relatedness, making them struggle to capture subtle inter-emotion distinctions.

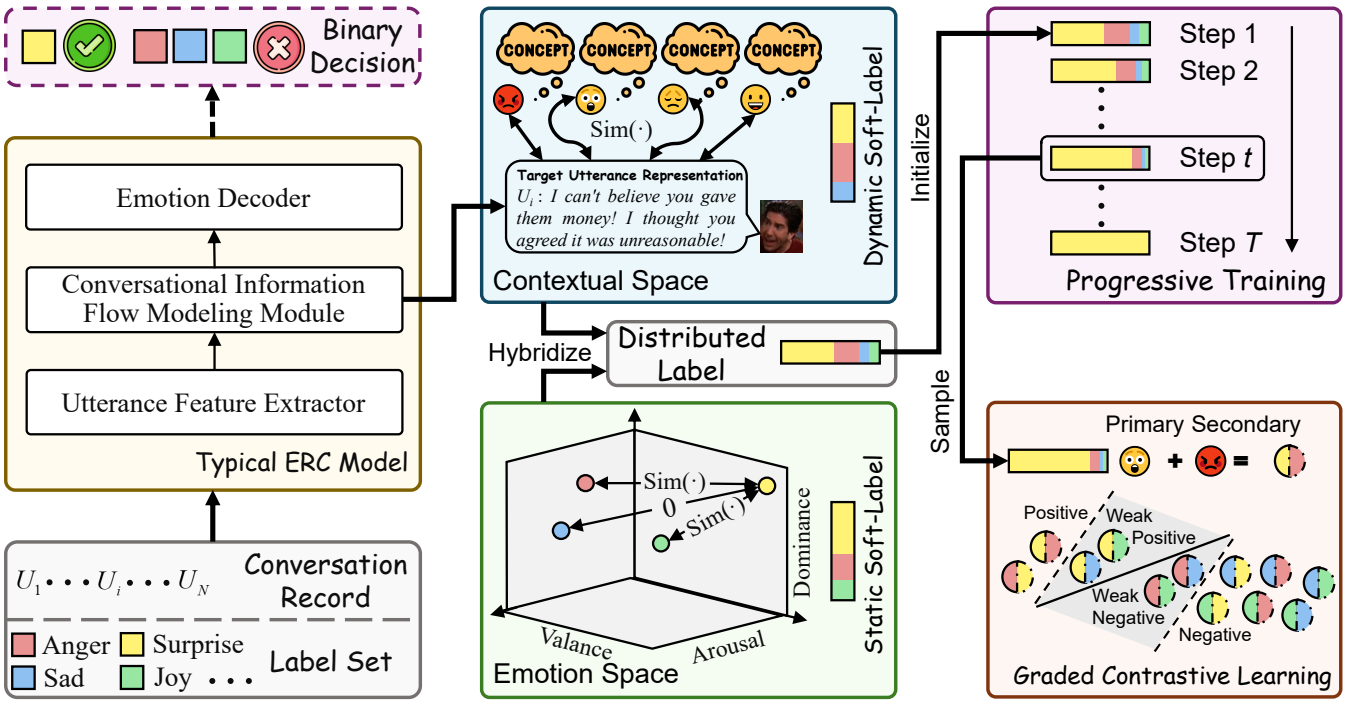

Figure 2: Overview of the proposed PHD framework. Sim(·) denotes the similarity function.

3 Proposed Framework

3.1 Problem Formulation

A dialogue can be formally represented as an ordered sequence of utterances $\langle u_1, ..., u_i, ..., u_N \rangle$, where $i \in [1, N]$ denotes the utterance index and N is the total number of utterances in the dialogue. Each utterance u_i is associated with its interlocutor $I_i \in \mathcal{I}$, where $|\mathcal{I}| \geq 2$. We further define an interlocutor–utterance pair as $U_i = (I_i, u_i)$, and represent the entire conversation as $\mathcal{C} = \{U_1, ..., U_N\}$. The goal of ERC is to recognize the discrete emotional state $y_i \in \mathcal{E}$ of the target utterance u_i conditioned on the conversational context, where $\mathcal{E}$ denotes the predefined emotion label set.

3.2 Typical ERC Pipeline

Typical ERC models follow a three-step sequential pipeline: Utterance Feature Extraction (UFE), Conversational Information Flow Modeling (CIFM), and Emotion Decoding (ED).

UFE first extracts context-independent features $\boldsymbol{u}_i$ for u_i:

$$\boldsymbol{u}_i = \mathrm{UFE}\left(u_i\right). \tag{1}$$

The CIFM module then performs conversational modeling over preceding context $u_{1:i-1}$ to capture temporal patterns under real-time settings, yielding a context-dependent emotion-aware hidden state $\boldsymbol{h}_i \in \mathbb{R}^d$ for each utterance u_i:

$$\boldsymbol{h}_i = \mathrm{CIFM}\left(\mathcal{C}, \boldsymbol{u}_i\right). \tag{2}$$

Finally, a emotion decoder with learnable parameters $\boldsymbol{W} \in \mathbb{R}^{|\mathcal{Y}| \times d}$ and $\boldsymbol{b} \in \mathbb{R}^{|\mathcal{Y}|}$ recognizes emotion probabilities of u_i:

$$\hat{\boldsymbol{y}}_i = \mathrm{softmax}\left(\boldsymbol{W}\boldsymbol{h}_i + \boldsymbol{b}\right). \tag{3}$$

Equations (1)-(3) outline the complete ERC process:

$$\hat{\boldsymbol{y}}_i = \mathrm{ERC}\left(\mathcal{C}, u_i\right). \tag{4}$$

Parameters are optimized via cross-entropy minimization:

$$L_{CE} = -\sum_{i=1}^{N} \sum_{e=1}^{|\mathcal{Y}|} y_{i,e} \log \hat{y}_{i,e}, \tag{5}$$

where $y_{i,e}$ and $\hat{y}_{i,e}$ denote the ground-truth and predicted probabilities for emotion e in utterance u_i, respectively.

3.3 Hybrid-Distributed Label Allocation

One-hot labeling constrains $y_{i,e}$ to a binary set, which limits its models to perceiving only emotion-shift triggers. We reconstruct emotion labels into distributed representations from both the emotion space and the contextual space, enabling the model not only to sense a shift but also to infer its tendency.

Static Emotion Space Modeling

Our modeling in the emotion space is grounded in the VAD theory [Mehrabian and Russell, 1974]. VAD posits that emotions are distributed in a continuous space with intrinsic coupling along three dimensions: *Valence*, *Arousal*, and *Dominance*. By representing emotions as continuous points in this space, the VAD framework provides a cognitively grounded perspective for characterizing inter-emotion relationships.

Datasets annotated with y_i's corresponding VAD representations $\boldsymbol{v}_i$ are rare and costly to scale. Without loss of generality, we extract VAD representations $\boldsymbol{v} \in \mathbb{R}^3$ for each emotion using the NRC-VAD lexicon [Mohammad, 2025].

We then compute the cosine similarity between $\boldsymbol{v}_i$ and the VAD representation $\boldsymbol{v}_e$ of each emotion $e \in \mathcal{E}$. The underlying relatedness between emotions is quantified as follows:

$$\mathrm{Sim}\,(i,e) = \begin{cases} \cos(\theta_{ie}) & \text{if } \boldsymbol{v}_i \cdot \boldsymbol{v}_e \geq 0, \\ 0 & \text{if } \boldsymbol{v}_i \cdot \boldsymbol{v}_e < 0, \end{cases} \tag{6}$$

where θ_{ie} denotes the angle between the two vectors, $\mathrm{Sim}\,(i,e)$ serves as a similarity score derived from the cosine similarity function $\cos(\cdot)$ and constitutes a component of the distributed emotion label.

We apply L1 normalization over all similarity components to construct the static distributional representation $\boldsymbol{y}_i^{static}$, ensuring its scale's consistency with the one-hot labeling:

$$y_{i,e}^{static} = \frac{\mathrm{Sim}\,(i,e)}{\sum_{e \in \mathcal{E}} \mathrm{Sim}\,(i,e)}, \tag{7}$$

where $y_{i,e}^{static}$ denotes the distribution component corresponding to emotion e. However, modeling emotions solely in the static space remains insufficient for capturing the subtle variations among utterances that share the same emotion label.

Dynamic Contextual Space Modeling

In the contextual space, we leverage $\boldsymbol{h}_i$ as the basis for allocating label components. The underlying intuition is that utterances sharing the same label may still exhibit subtle contextual semantic variations, which are often critical in shaping the direction of subsequent emotion shifts.

To capture these nuances, we first retrieve the top-C confidence concepts associated with each emotion e from ConceptNet [Speer *et al.*, 2017]. These concepts are concatenated with e itself and used to replace $\boldsymbol{u}_i$. Leveraging the information-filtering capability of the CIFM module, irrelevant concepts are suppressed, yielding a context-aware emotion representation $\boldsymbol{h}_e$ that resides in the same space as $\boldsymbol{h}_i$.

Following the same procedure as in Equations (6)-(7), we then compute the cosine similarity between $\boldsymbol{h}_i$ and $\boldsymbol{h}_e$ to obtain the component of the dynamic distributed label. These components are then proportionally normalized to form the dynamic distributional representation $\boldsymbol{y}_i^{dynamic}$.

Finally, we introduce a tunable hyperparameter $\lambda \in [0,1]$ to control the fusion of the static and dynamic components, resulting in the final hybrid-distributional label $\boldsymbol{y}_i^{dis}$:

$$\boldsymbol{y}_i^{dis} = \lambda \boldsymbol{y}_i^{static} + (1-\lambda)\boldsymbol{y}_i^{dynamic}. \tag{8}$$

Compared to the original one-hot label $\boldsymbol{y}_i$, the reconstructed soft label $\boldsymbol{y}_i^{dis}$ not only encodes the latent relatedness among emotions but also captures the multiple affective tendencies inherent to the utterance within its specific context.

3.4 Progressive Training Strategy

The conventional training paradigm for ERC, as formulated in Equation (5), aims to match each utterance with its most appropriate emotion label. However, human emotional expression during conversation is multi-faceted and entangled, and identifying only the dominant emotion inevitably discards other informative affective cues. Leveraging the unique advantages of the reconstructed distributional labels $\boldsymbol{y}_i^{dis}$, we further propose a novel Progressive Training Strategy (PTS).

Specifically, at the early stage, we treat $\boldsymbol{y}_i^{dis}$ as the target probabilities. That is, each utterance is not only aligned with its ground-truth label but is also allowed to partially map to other emotions. At each training step, we smoothly guide the model from capturing mixed emotional states toward the one-hot matching according to the following update rules:

$$y_{i,e}^{t+1} = \begin{cases} \frac{1}{1+\varepsilon \sum_{e \neq y_i} y_{i,e}^t} & \text{if } e = y_i, \\ \frac{\varepsilon y_{i,e}^t}{1+\varepsilon \sum_{e \neq y_i} y_{i,e}^t} & \text{if } e \neq y_i, \end{cases} \tag{9}$$

$$y_{i,e}^0 = y_{i,e}^{dis}, \tag{10}$$

$$\varepsilon = (T-t)/(T-1), \tag{11}$$

where $(t+1) \in [1,T]$ denotes the current step, T is the maximum number of training steps, and y_i represents the ground-truth emotion for u_i. The coefficient ε decays linearly from 1 to 0 during training. Equation (9) ensures y_i consistently maintains its dominant proportion within the evolving label.

At each training step, the updated label distribution $\boldsymbol{y}_i^{t+1}$ serves as the new supervision signal, replacing Equation (5) with the following Kullback-Leibler divergence objective:

$$L_{KL} = -\sum_{i=1}^{N} \sum_{e=1}^{|\mathcal{E}|} \sum_{t=0}^{T-1} \boldsymbol{y}_i^{t+1} \log \frac{\hat{\boldsymbol{y}}_i}{\boldsymbol{y}_i^{t+1}}, \tag{12}$$

where the scheduling in Equation (11) guarantees that the training target gradually converges to one-hot supervision.

Theoretically, PTS *performs a gradual entropy contraction of the supervision signal*. This process is mathematically isomorphic to interpolation-based distillation, which has been shown to stabilize training in non-convex optimization settings [Shing *et al.*, 2025].

3.5 Graded Contrastive Learning

Supervised Contrastive Learning (SCL) cannot be directly applied in the distributional setting, since each soft label is theoretically unique and thus cannot be grouped under conventional class-based definitions. Moreover, rigid binary partitioning rule inevitably leads to intra-set conflicts, where certain negative samples are closer to the anchor than some positives. In addition, the long-tailed distribution further exacerbates the imbalance between positive and negative sets.

Inspired by graded evaluation systems [Kang and Cho, 2025], we design a Graded Contrastive Learning mechanism to address these issues. Beyond binary partition, GCL further introduces two intermediate sets, weak positives and weak negatives. Concretely, we assign u_i a primary emotion y_i^p and a secondary emotion y_i^s according to the component-wise ranking based on $\boldsymbol{y}_i^{dis}$. For each sample j in the same mini-batch, we define the following sets with respect to anchor i:

$$P(i) = \left\{ \forall j \in U, \left\{ y_j^p, y_j^s \right\} = \left\{ y_i^p, y_i^s \right\} \right\}, \tag{13}$$

$$WP(i) = \left\{ \forall j \in U, y_j^p = y_i^p \cap y_j^s \neq y_i^s \right\}, \tag{14}$$

$$WN(i) = \left\{ \forall j \in U \backslash P(i), y_j^p = y_i^s \right\}, \tag{15}$$

$$N(i) = \left\{ \forall j \in U \backslash \left\{ P(i) \cup WP(i) \cup WP(i) \right\} \right\}, \tag{16}$$

where $P(i)$, $N(i)$, $WP(i)$, and $WN(i)$ denote the positive, negative, weak-positive, and weak-negative sets, respectively, and U represents the set of all samples in the batch.

Algorithm 1 Training Process with PHD

Input: Conversation set $\mathcal{C}$, emotion set $\mathcal{E}$, max training steps T, fusion weight λ, temperature τ, GCL weight α
Output: Trained ERC model parameters Θ

Initialize model parameters Θ
Precompute emotion-level VAD embeddings $\{\boldsymbol{v}_e\}_{e\in\mathcal{E}}$
for each training step $t+1=1$ to T **do**
  Sample a mini-batch $\mathcal{B}\subset\mathcal{C}$
  $\varepsilon \leftarrow \frac{T-t}{T-1}$ {PTS decay factor}
  for each utterance $u_i \in \mathcal{B}$ **do**
    $\boldsymbol{u}_i, \boldsymbol{h}_i, \hat{\boldsymbol{y}}_i \leftarrow \mathrm{ERC}(\mathcal{C}, u_i)$
    $\boldsymbol{y}_i^{static} \leftarrow \mathrm{StaticAlloc}(\boldsymbol{v}_{y_i}, \{\boldsymbol{v}_e\})$
    $\boldsymbol{y}_i^{dynamic} \leftarrow \mathrm{DynamicAlloc}(\boldsymbol{h}_i, \{\boldsymbol{h}_e\})$
    $\mathbf{y}_i^{dis} \leftarrow \lambda \mathbf{y}_i^{static} + (1-\lambda)\mathbf{y}_i^{dynamic}$
    $\mathbf{y}_i^{t+1} \leftarrow \mathrm{PTS\_update}(\mathbf{y}_i^t, y_i, \varepsilon)$
  end for
  $L_{KL} \leftarrow \mathrm{KL}(\{\boldsymbol{y}_i^{t+1}\}, \hat{\boldsymbol{y}}_i)$
  Construct graded sets $\{P(i), WP(i), WN(i), N(i)\}$ for each $i \in \mathcal{B}$
  $L_{GCL} \leftarrow \mathrm{GCL\_loss}(\boldsymbol{h}_i, \boldsymbol{y}_i^{dis}, \tau)$
  $L_{PHD} \leftarrow L_{KL} + \alpha L_{GCL}$
  Update Θ by minimizing L_{PHD}
end for

The composition-based criterion in Equation (13) and the primary-emphasis rule in Equation (14) resolve the sample conflict issue. The GCL part in Figure 2 enumerates all possible partitions under four emotion scenarios, achieving a more balanced distribution among sets, thereby alleviating the problem of insufficient positive samples.

To regulate attraction and repulsion across sets, we assign graded weights, which is formalized as follows:

$$\mathcal{P}(i) = \sum_{j\in P(i)} \exp\left(\cos\left(\boldsymbol{h}_i, \boldsymbol{h}_j\right)/\tau\right) + \\ + \frac{\cos\left(\boldsymbol{y}_i^{dis}, \boldsymbol{y}_j^{dis}\right)}{2} \sum_{j\in WP(i)} \exp\left(\cos\left(\boldsymbol{h}_i, \boldsymbol{h}_j\right)/\tau\right), \quad (17)$$

$$\mathcal{N}(i) = \sum_{j\in N(i)} \exp\left(\cos\left(\boldsymbol{h}_i, \boldsymbol{h}_j\right)/\tau\right) + \\ + \frac{1-\cos\left(\boldsymbol{y}_i^{dis}, \boldsymbol{y}_j^{dis}\right)}{2} \sum_{j\in WN(i)} \exp\left(\cos\left(\boldsymbol{h}_i, \boldsymbol{h}_j\right)/\tau\right), \quad (18)$$

where τ denotes the temperature parameter. The resulting GCL loss for a mini-batch $\mathcal{B}$ is then defined as:

$$L_{GCL} = -\sum_{i=1}^{N}\sum_{j=1}^{|\mathcal{B}|} \frac{1}{|P(i)|}\frac{\mathcal{P}(i)}{\mathcal{N}(i)}. \quad (19)$$

The total training objective of our PHD framework consists of the progressive cross-entropy loss and the GCL loss:

$$L_{PHD} = L_{KL} + \alpha L_{GCL}, \quad (20)$$

where α is a hyperparameter controlling the contribution of L_{GCL}. The complete optimization procedure of the PHD framework is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Settings

Datasets and Metrics

We evaluate the proposed PHD framework on three ERC benchmark datasets: IEMOCAP [Busso *et al.*, 2008], MELD [Poria *et al.*, 2019], and EmoryNLP [Zahiri and Choi, 2018]. Following standard practice, performance is measured using the weighted F1 score (w-F1).

Baselines

To ensure a comprehensive comparison, we integrate PHD with five representative baseline methods covering diverse technical paradigms: GNN-based **DualGATs** [Zhang *et al.*, 2023], knowledge-enhanced **MKFM** [Tu *et al.*, 2023b], prompt-based **ERC-DP** [Wang *et al.*, 2024], and LLM-based **DialogueLLM** [Zhang *et al.*, 2025] and **CoE** [Shen *et al.*, 2025b]. Detailed descriptions are provided in Section 2.1.

Implementation Details

For fair comparison, we retain the original configurations of all backbone models. Each backbone is first fine-tuned for several epochs using its standard training procedure and is then incorporated into the PHD framework for further optimization. For LLM-based methods, we introduce the following adaptations to ensure compatibility with PHD:

(A) Instruction Template Adaptation. The instruction template is modified to end with: "*Analyze the interlocutor's emotional state based on the dialogue. Output a probability distribution over the predefined emotion categories. The probabilities should sum to 1.*" Correspondingly, the structured output format is adjusted to align with $\boldsymbol{y}_i^{dis}$.

(B) Dynamic Label Construction. For each utterance, we fix the prompt as: "*Given the dialogue context and the current utterance, estimate the interlocutor's emotional tendencies as a probability distribution.*" The resulting LLM outputs are used to construct the dynamic distribution $\boldsymbol{y}^{dynamic}$.

(C) GCL Adaptation for LLMs. We adapt GCL to LLMs by operating on the logits produced by the LLM's decoder. Specifically, for each training instance u_i, the model produces a emotion-related logit vector $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{E}|}$. Based on $y_{i,e}^{dis}$, we derive ordered emotion pairs (e_m, e_n) whenever $y_{i,m}^{dis} > y_{i,n}^{dis}$. The supervision strength of each pair is weighted by the probability gap $\Delta_{mn} = y_{i,m}^{dis} - y_{i,n}^{dis}$. L_{GCL} is thus reformulated as a weighted pairwise ranking loss:

$$L_{GCL} = \sum_{(m,n)} \Delta_{mn} \cdot \log\left(1+\exp(z_n - z_m)\right), \quad (21)$$

which enables graded supervision for mixed emotional states, avoids rigid binary partitioning, and can be seamlessly integrated into standard instruction tuning via back propagation.

The temperature parameter $\tau = 0.07$. The fusion weight λ is set to 0.3 on IEMOCAP and 0.4 on the other datasets, while L_{GCL} weight $\alpha = 0.15$. All experiments are conducted on two NVIDIA RTX A6000 GPUs. Reported results are averaged over five runs with random initialization. A paired t-test with a significance level of 0.05 is performed to validate the statistical significance of performance improvements.

Methods	IEMOCAP	MELD	EmoryNLP	Average
DualGATs	67.68	66.90	40.69	58.42
w/ PHD	69.70	68.52	41.24	59.82
Improvement	↑2.02	↑1.62	↑0.55	↑**1.40**
MKFM	68.08	65.50	39.76	57.78
w/ PHD	69.85	67.85	40.65	59.45
Improvement	↑1.77	↑2.35	↑0.89	↑**1.67**
ERC-DP	69.64	67.34	40.10	59.03
w/ PHD	70.72	68.60	40.92	60.08
Improvement	↑1.08	↑1.26	↑0.82	↑**1.05**
DialogueLLM*	69.87	68.54	40.28	59.56
w/ PHD	70.94	70.11	41.12	60.72
Improvement	↑1.07	↑1.57	↑0.84	↑**1.16**
CoE*	70.12	67.96	40.81	59.63
w/ PHD	71.14	69.62	41.62	60.79
Improvement	↑1.02	↑1.66	↑0.81	↑**1.16**

Table 1: Overall performance comparison of backbone models before and after integrating PHD. * denotes reproduced results.

Methods	DualGATs	DialogueLLM	CoE
PHD	59.82	60.72	60.79
w/o static	59.42(↓0.40)	59.98(↓0.74)	60.08(↓0.71)
w/o dynamic	58.30(↓1.52)	58.85(↓1.87)	58.93(↓1.86)
w/o PTS	57.54(↓2.28)	58.69(↓2.03)	58.87(↓1.92)
w/o GCL	58.71(↓1.11)	59.84(↓0.88)	59.84(↓0.95)

Table 2: Ablation results (average performance on three datasets).

4.2 Performance Comparison

As reported in Table 1, PHD consistently yields significant performance gains, demonstrating strong architectural portability. On average, PHD improves overall performance by 1.29 w-F1 points, with gains ranging from 1.05 to 1.67.

Notably, larger gains are observed on IEMOCAP and MELD, which are known to contain a higher proportion of emotion-shift utterances and neutral-valence transitions. The improvement on EmoryNLP is relatively moderate, suggesting that PHD does not introduce bias when emotional dynamics are relatively stable, but instead behaves conservatively.

The effectiveness of PHD further extends to LLM-based approaches, indicating that even LLMs remain constrained by one-hot emotion supervision. By injecting graded emotion relevance and progressive entropy contraction, PHD complements LLM reasoning, leading to additional improvements.

Overall, these results demonstrate that PHD functions as a model-agnostic, plug-and-play training framework. More importantly, they confirm that modeling emotional dynamics as a progressive, distributional process is critical for ERC.

4.3 Ablation Study

We conduct ablation experiments to quantify the contribution of each component and to validate their complementary roles. Specifically, *w/o static* removes the emotion-space distribution derived from VAD, *w/o dynamic* discards the context-aware distribution allocated in the conversational space, *w/o PTS* disables the progressive contraction strategy and trains with fixed labels, and *w/o GCL* removes L_{GCL} from L_{PHD}.

Methods	IEMOCAP	MELD	EmoryNLP	Average
DualGATs	59.30	57.49	36.28	51.02
w/ PHD	64.53	62.83	37.98	55.11
Improvement	↑5.23	↑5.34	↑1.70	↑**4.09**
MKFM	59.68	57.02	36.21	50.97
w/ PHD	64.34	60.51	38.38	54.41
Improvement	↑4.66	↑3.49	↑2.17	↑**3.44**
ERC-DP	56.66	51.50	35.41	47.86
w/ PHD	61.36	57.39	39.89	52.88
Improvement	↑4.70	↑5.89	↑2.48	↑**4.36**
DialogueLLM	58.76	60.16	37.13	52.02
w/ PHD	61.91	64.40	39.28	55.20
Improvement	↑3.15	↑4.24	↑2.15	↑**3.18**
CoE	60.08	59.35	37.82	52.42
w/ PHD	62.33	62.44	39.54	54.77
Improvement	↑2.25	↑3.09	↑2.41	↑**2.58**

Table 3: Performance comparison of backbone models before and after integrating PHD in emotion-shift scenarios.

As shown in Table 2, removing any component leads to consistent performance degradation, confirming that each module contributes meaningfully to PHD. Removing the dynamic component causes more pronounced drop than the static, highlighting the importance of context-dependent distributional cues for modeling emotional tendencies beyond label semantics. Disabling PTS results in the largest overall degradation, demonstrating that progressive supervision is critical for stabilizing training and mimicking gradual emotional evolution. The performance decrease observed without GCL verifies the necessity of graded sample partitioning for mitigating label conflicts under distributional supervision.

4.4 Performance on Emotion Shift

We further evaluate PHD specifically on emotion-shift scenarios to examine its ability to capture emotional transitions. As reported in Table 3, PHD consistently outperforms backbones across all emotion-shift subsets. Notably, the gains are more pronounced on MELD, where frequent subtle shifts occur, highlighting PHD's strength in anticipating the direction of emotion changes rather than merely detecting shifts.

Overall, these findings validate that PHD significantly enhances the model's sensitivity to evolving emotional dynamics, addressing a key limitation of existing methods.

4.5 Performance on Neutral-Valence Emotions

We further investigate the effectiveness of PHD on neutral-valence-initiated emotion transitions, which are particularly challenging for conventional ERC models due to their subtlety and low signal in one-hot supervision. As reported in Table 4, integrating PHD substantially improves performance across all datasets. The improvements are particularly pronounced for subtle valence shifts, such as Neutral-to-Other in IEMOCAP and Peaceful-to-Other in EmoryNLP. These results demonstrate that PHD effectively enhances ERC models' sensitivity to neutral-valence dynamics, enabling more reliable recognition of gradual emotional transitions.

Methods	IEMOCAP	MELD		EmoryNLP	
	Neutral→ $\mathcal{E}$\Neutral	Neutral→ $\mathcal{E}$\Neutral	Surprise→ $\mathcal{E}$\Surprise	Neutral→ $\mathcal{E}$\Neutral	Peaceful→ $\mathcal{E}$\Peaceful
DualGATs	34.12	41.32	31.89	23.90	24.00
w/ PHD	47.00 (↑12.88)	54.92 (↑13.60)	38.22 (↑6.33)	30.16 (↑6.26)	34.53 (↑10.53)
DialogueLLM	38.14	36.95	38.82	31.31	28.36
w/ PHD	52.00 (↑13.86)	43.31 (↑6.36)	44.19 (↑5.37)	43.15 (↑11.84)	35.47 (↑7.11)
CoE	43.37	37.10	37.73	29.72	28.03
w/ PHD	55.24 (↑11.87)	44.54 (↑7.44)	47.31 (↑9.58)	39.36 (↑9.64)	35.82 (↑7.79)

Table 4: Performance comparison of backbone models before and after integrating PHD in neutral-valence-initiated emotion-shift scenarios.

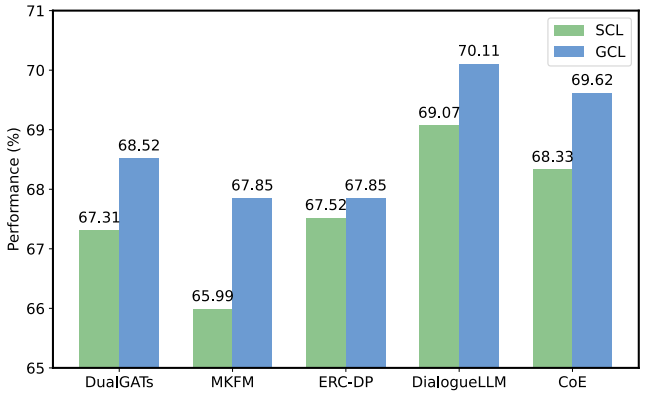

Figure 3: Performance comparison under different contrastive mechanisms on the MELD dataset.

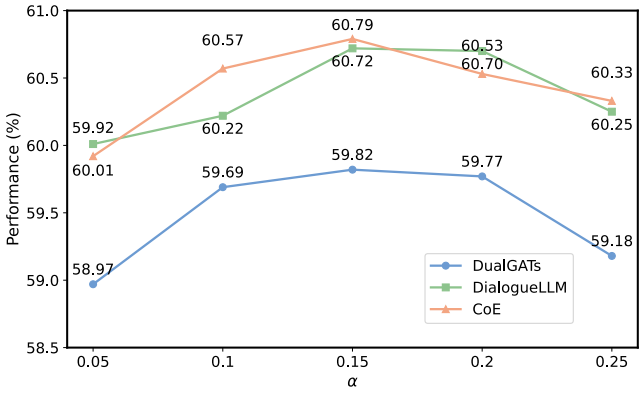

Figure 4: Performance with varying L_{GCL} weight α.

4.6 Effect of Graded Contrastive Learning

Qualitative Analysis. We further conduct a qualitative comparison between GCL and conventional SCL. As illustrated in Figure 3, GCL consistently outperforms SCL across all backbone models. The largest gain is observed on MKFM, demonstrating that GCL can effectively leverage nuanced inter-sample relationships to enhance feature representation. These results highlight the superiority of GCL over conventional SCL, indicating that modeling graded sample similarities can lead to more discriminative embeddings.

Quantitative Analysis. We conduct experiments by varying L_{GCL} weight α from 0.05 to 0.25 to determine the optimal balance between the GCL loss and the primary task loss. As summarized in Figure 4, the performance initially improves as α increases from 0.05 to 0.15. Beyond this point, further increasing leads to a slight decline. The trends indicates that a moderate GCL weight effectively enhances the embedding quality while maintaining task-specific discriminability. Overall, a moderate α maximizes the benefits of graded contrastive supervision, yielding more discriminative embeddings and superior model performance.

4.7 Computational Efficiency

We evaluate the training efficiency of PHD. As shown in Table 5, for DualGATs, the additional cost remains within 5.2%-9.6% range, while for DialogueLLM, the relative increase is slightly higher but still stable. Importantly, the absolute training time is largely dominated by the backbone itself.

Methods	IEMOCAP	MELD	EmoryNLP
DualGATs	52	82	77
w/ PHD	57(↑9.6%)	89(↑8.5%)	81(↑5.2%)
DialogueLLM	153	320	289
w/ PHD	171(↑11.8%)	359(↑12.2%)	323(↑11.8%)

Table 5: Training time(minute) comparison of backbone models before and after integrating PHD.

5 Conclusion

In this work, we introduce a novel, plug-and-play learning framework PHD for modeling emotion transitions that addresses the inherent limitations of one-hot supervision. By leveraging hybrid-distributional labels, a progressive training strategy, and graded contrastive learning, PHD enables models to capture both gradual emotion transitions and the entangled nature of affective states. Extensive evaluations across multiple benchmarks demonstrate that PHD consistently enhances performance of both conventional and LLM-based architectures, particularly in emotion-shift and neutral-valence scenarios. Ablation analyses confirm that each component contributes uniquely to the overall improvement, highlighting the synergy between static emotion-space modeling, dynamic context-aware allocation, progressive supervision, and graded contrastive reasoning. Our findings underscore that incorporating distributional and progressive perspectives is pivotal for building systems capable of robust and fine-grained emotional understanding in real-world dialogues.

Acknowledgements

We thank the reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62276059 and by the Fundamental Research Funds for the Central Universities under Grants No. 2572025JT05 and 2572026BR15.

References

[Baidu, 2023] Baidu. Ernie bot, 2023. Chatbot, based on the ERNIE large-language-model series. Released March 16, 2023.

[Barrett, 2017] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social cognitive and affective neuroscience*, 12(1):1–23, 2017.

[Busso *et al.*, 2008] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[Cambria *et al.*, 2026] Erik Cambria, Rui Mao, Xulang Zhang, Luwei Xiao, Tiesunlong Shen, and Avinash Anand. Senticnet 9: Generative commonsense for emotion ai via conceptual primitive discovery and time shift mechanism. *IEEE Transactions on Computational Social Systems*, 2026.

[Cao *et al.*, 2025] YuKun Cao, Luobin Huang, and Yijia Tang. Petracker: Poincaré-based dual-strategy emotion tracker for emotion recognition in conversation. *IEEE Transactions on Affective Computing*, 16(3):2020–2032, 2025.

[Dai *et al.*, 2024] Yijing Dai, Yingjian Li, Dongpeng Chen, Jinxing Li, and Guangming Lu. Multimodal decoupled distillation graph neural network for emotion recognition in conversation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9910–9924, 2024.

[Hu *et al.*, 2021] Dou Hu, Lingwei Wei, and Xiaoyong Huai. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052, 2021.

[Ji *et al.*, 2025] Hongru Ji, Xianghua Li, Mingxin Li, Meng Zhao, and Chao Gao. Hybrid relational graphs with sentiment-laden semantic alignment for multimodal emotion recognition in conversation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 2973–2981, 2025.

[Joshi *et al.*, 2022] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. Cogmen: Contextualized gnn based multimodal emotion recognition. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4148–4164, 2022.

[Kang and Cho, 2024] Yujin Kang and Yoon-Sik Cho. Improving contrastive learning in emotion recognition in conversation via data augmentation and decoupled neutral emotion. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2194–2208, 2024.

[Kang and Cho, 2025] Yujin Kang and Yoon-Sik Cho. Beyond single emotion: Multi-label approach to conversational emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24321–24329, 2025.

[Kuppens *et al.*, 2010] Peter Kuppens, Nicholas B Allen, and Lisa B Sheeber. Emotional inertia and psychological maladjustment. *Psychological science*, 21(7):984–991, 2010.

[Li *et al.*, 2025] Jiang Li, Xiaoping Wang, and Zhigang Zeng. Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8786–8803, 2025.

[Liu *et al.*, 2022] Yuchen Liu, Jinming Zhao, Jingwen Hu, Ruichen Li, and Qin Jin. Dialogueein: Emotion interaction network for dialogue affective analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 684–693, 2022.

[Mao *et al.*, 2023] Yuzhao Mao, Di Lu, Yang Zhang, and Xiaojie Wang. Fatrer: Full-attention topic regularizer for accurate and robust conversational emotion recognition. In *ECAI 2023*, pages 1688–1695. IOS Press, 2023.

[Mao *et al.*, 2025] Rui Mao, Guanyi Chen, Xiao Li, Mengshi Ge, and Erik Cambria. A comparative analysis of metaphorical cognition in chatgpt and human minds. *Cognitive Computation*, 17(1):35, 2025.

[Mehrabian and Russell, 1974] A. Mehrabian and J.A. Russell. *An Approach to Environmental Psychology*. M.I.T. Press, 1974.

[Mohammad, 2025] Saif M Mohammad. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *arXiv preprint arXiv:2503.23547*, 2025.

[Nie *et al.*, 2023] Weizhi Nie, Yuru Bao, Yue Zhao, and Anan Liu. Long dialogue emotion detection based on commonsense knowledge graph guidance. *IEEE Transactions on Multimedia*, 26:514–528, 2023.

[Pereira *et al.*, 2025] Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. Deep emotion recognition in textual conversations: A survey. *Artificial Intelligence Review*, 58(1):10, 2025.

[Poria *et al.*, 2019] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada 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*, pages 527–536, 2019.

[Roohi *et al.*, 2025] Samad Roohi, Richard Skarbez, and Hien Duy Nguyen. Reliable uncertainty estimation in emotion recognition in conversation using conformal prediction framework. *Natural Language Processing*, 31(5):1163–1186, 2025.

[Shen *et al.*, 2025a] Siyuan Shen, Feng Liu, Hanyang Wang, and Aimin Zhou. Towards speaker-unknown emotion recognition in conversation via progressive contrastive deep supervision. *IEEE Transactions on Affective Computing*, 16(3):2261–2273, 2025.

[Shen *et al.*, 2025b] Zhiyu Shen, Yunhe Pang, Yanghui Rao, and Jianxing Yu. Coe: A clue of emotion framework for emotion recognition in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23548–23563, 2025.

[Shing *et al.*, 2025] Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. TAID: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Speer *et al.*, 2017] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[Sun *et al.*, 2021] Yang Sun, Nan Yu, and Guohong Fu. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2949–2958, 2021.

[Tu *et al.*, 2023a] Geng Tu, Bin Liang, Ruibin Mao, Min Yang, and Ruifeng Xu. Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations. In *Findings of the association for computational linguistics: ACL 2023*, pages 14054–14067, 2023.

[Tu *et al.*, 2023b] Geng Tu, Bin Liang, Bing Qin, Kam-Fai Wong, and Ruifeng Xu. An empirical study on multiple knowledge from chatgpt for emotion recognition in conversations. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 12160–12173, 2023.

[Tu *et al.*, 2024] Geng Tu, Feng Xiong, Bin Liang, and Ruifeng Xu. A persona-infused cross-task graph network for multimodal emotion recognition with emotion shift detection in conversations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2266–2270, 2024.

[Wang *et al.*, 2024] Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. Emotion recognition in conversation via dynamic personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5711–5722, 2024.

[Xie *et al.*, 2025] Yunhe Xie, Cheng-Jie Sun, Ziyi Cao, Bingquan Liu, Zhenzhou Ji, Yuanchao Liu, and Lili Shan. A dual contrastive learning framework for enhanced multimodal conversational emotion recognition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4055–4065, 2025.

[Xu *et al.*, 2025] Qinfu Xu, Yiwei Wei, Chunlei Wu, Leiquan Wang, Shaozu Yuan, Jie Wu, Jing Lu, and Hengyang Zhou. Towards multimodal sentiment analysis via hierarchical correlation modeling with semantic distribution constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21788–21796, 2025.

[Yang *et al.*, 2022] Lin Yang, Yi Shen, Yue Mao, and Longjun Cai. Hybrid curriculum learning for emotion recognition in conversation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11595–11603, 2022.

[Yang *et al.*, 2025] Juan Yang, Puling Wei, Xu Du, and Jun Shen. Graph attention based on contextual reasoning and emotion-shift awareness for emotion recognition in conversations. *Complex & Intelligent Systems*, 11(7):287, 2025.

[Zahiri and Choi, 2018] Sayyed M Zahiri and Jinho D Choi. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*, 2018.

[Zha *et al.*, 2025] Xupeng Zha, Huan Zhao, Guanghui Ye, and Zixing Zhang. Dual-view learning for conversational emotion recognition through context and emotion-shift modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25823–25831, 2025.

[Zhang and Li, 2023] Xiaoheng Zhang and Yang Li. A cross-modality context fusion and semantic refinement network for emotion recognition in conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13099–13110, 2023.

[Zhang *et al.*, 2023] Duzhen Zhang, Feilong Chen, and Xiuyi Chen. Dualgats: Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7395–7408, 2023.

[Zhang *et al.*, 2025] Yazhou Zhang, Mengyao Wang, Youxi Wu, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. *Neural Networks*, page 107901, 2025.

[Zhao *et al.*, 2022] Weixiang Zhao, Yanyan Zhao, and Bing Qin. Mucdn: Mutual conversational detachment network for emotion recognition in multi-party conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7020–7030, 2022.