

Article

Task-Adaptive and Multi-Level Contextual Understanding for Emotion Recognition in Conversations

Xiaomeng Yao ^{1,2,*}, Wei Cao ^{1,2,*}, Yuyang Xue ³, Haijun Zhang ^{1,2}  and Xiaochao Fan ^{1,2}

¹ School of Computer Science and Technology, Xinjiang Normal University, Urumqi 830054, China; yxm1428@163.com (X.Y.); fxc1982@xjnu.edu.cn (X.F.)

² Xinjiang Engineering Research Center for Smart Education and Application, Urumqi 830054, China

³ The School of Information Management, Xinjiang University of Finance & Economics, Urumqi 830012, China; xyy@xjufe.edu.cn

* Correspondence: cw0808@mail.ustc.edu.cn

Abstract

Emotion recognition in conversations (ERC) is a significant task in natural language processing, aimed at identifying the emotion of each utterance within a conversation. Current research predominantly relies on pre-trained language models, often incorporating sophisticated network architectures to capture complex contextual semantics in conversations. However, existing approaches have not successfully combined effective task-specific adaptation with adequate modeling of conversational context complexity. To address this, we propose a model named TAMC-ERC (Task-Adaptive and Multi-level Contextual Understanding for Emotion Recognition in Conversations). The model adopts a progressive recognition framework that sequentially builds on foundational utterance representations, integrates conversation-level contexts, and leads to a task-adaptive classification decision. First, the Task-Adaptive Representation Learning module produces highly discriminative utterance representations. It achieves this by integrating emotion space information into prompts and employing contrastive learning. Subsequently, the Multi-Level Contextual Understanding module performs in-depth modeling of the conversational context. It synergistically integrates both macroscopic narratives and microscopic interactions to construct a comprehensive emotional context. Finally, the classifier is directly parameterized by the emotion concept vectors from the task-adaptive stage. This creates a coherent task adaptation process, maintaining task-specific awareness from representation learning through to the final decision. Experiments on three benchmark datasets demonstrate that TAMC-ERC achieves highly competitive performance: it attains weighted average F1 scores of 71.04% on IEMOCAP, 66.95% on MELD, and 40.99% on EmoryNLP. These results set a new state of the art and demonstrate that the model outperforms most existing baselines. This work validates that integrating task adaptation with multi-level contextual modeling is key to addressing conversational complexity and improving recognition accuracy.



Academic Editor: Douglas O'Shaughnessy

Received: 20 December 2025

Revised: 28 January 2026

Accepted: 4 February 2026

Published: 9 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: emotion recognition in conversations (ERC); task adaptation; contextual modeling; conversational understanding

1. Introduction

Emotion recognition in conversations (ERC) has emerged as a significant task in natural language processing, aiming to automatically identify the emotional state of each utterance within a conversation. This technology finds broad applications in intelligent customer service, social media analysis, and human–computer interaction.

Early research primarily focused on modeling temporal dependencies using recurrent neural networks, as seen in DialogueRNN [1]. To more effectively capture complex inter-utterance interactions, graph-structured models were subsequently developed. For instance, DialogueGCN [2] represents utterances as graph nodes, utilizing various edge types to model speaker relationships. COGMEN [3] and DAG-ERC [4] employed graph Transformers and directed acyclic graphs, respectively, to handle global/local semantics and spatiotemporal structures. Concurrently, advanced learning objectives emerged to improve utterance representations. Representative works include Cog-BART [5], which integrates contrastive learning with BART [6], and SACL [7], which employs adversarial training for more robust representations. With pre-trained language models (PLMs) becoming the cornerstone technology, how to effectively adapt them to capture conversational emotional dynamics has emerged as a new focus.

Despite considerable progress, existing approaches face two fundamental challenges. First, pre-trained language models (PLMs) possess no intrinsic knowledge of the task-specific emotion space. This fundamental gap prevents them from performing accurate, task-aware semantic reasoning for emotion recognition. Second, existing approaches struggle to balance the macroscopic narrative flow and microscopic interactions in conversations, leading to an incomplete and superficial construction of the emotional context. Therefore, these two challenges define a progressive research problem. The first challenge concerns the acquisition of high-quality foundational utterance representations, while the second addresses the refinement of these representations within the complete conversational context.

Accordingly, this work aims to develop a new framework that addresses two progressive challenges. The first is the adaptation of PLMs to the task's emotion space for obtaining high-quality foundational utterance representations. The second involves the design of a context encoder capable of comprehending and integrating both macroscopic narrative flow and microscopic interaction patterns to construct a complete and in-depth emotional context. Crucially, these two challenges are interlinked: high-quality task-adaptive representations provide a solid foundation for context modeling, while a comprehensively understood context delivers the necessary informational basis for final emotion classification. Therefore, a synergistic framework that addresses both in concert is required.

To this end, this paper proposes TAMC-ERC, advancing ERC research from both task adaptation and complex contextual understanding perspectives. Specifically, the model first processes utterances through a task-adaptive module that injects the dataset's emotion space into the PLM, yielding high-quality representations. A multi-level contextual module then synergistically integrates macroscopic narratives and microscopic interactions to build a comprehensive emotional context. Finally, a classifier, parameterized by emotion concept vectors from the task-adaptive stage, ensures systematic task adaptation from feature learning to the final decision.

The main contributions of this work are as follows:

- (1) A task-adaptive representation learning mechanism that enhances PLM's adaptability to specific emotion tasks through emotion space-constrained prompt learning and contrastive fine-tuning;
- (2) A multi-level contextual understanding architecture designed to systematically handle the macroscopic narrative flow and microscopic interactions in conversations;
- (3) Extensive experiments on benchmark datasets demonstrate that TAMC-ERC achieves highly competitive performance, with ablation studies confirming a synergistic effect between its core components.

2. Related Work

This section reviews previous research closely related to our work. We first summarize major approaches for context modeling in emotion recognition in conversations, including sequence-based, graph-based, and knowledge-fusion methods. We then review advances in contrastive learning, which is highly relevant to our proposed framework.

2.1. Emotion Recognition in Conversations

2.1.1. Sequence-Based Models

Early deep learning approaches for ERC predominantly leveraged sequential models. Majumder et al. [1] proposed DialogueRNN, which employs multiple RNNs to track dialogue states and speaker dynamics. Subsequent extensions include Cog-BART [5], which integrates a generative pre-trained model with contrastive learning, and DialogueEIN [8], which explicitly models emotional inertia and stimuli. To enhance robustness, Hu et al. [7] proposed an adversarial sample-learning framework (SACL), while Yang et al. [9] developed a modality-interaction framework (SCMM) to address semantic limitations. Furthermore, Yang et al. [10] explored a holistic curriculum-learning paradigm, and Lei et al. [11] and Zhang et al. [12] investigated instruction tuning for LLMs on the ERC task. Lee et al. [13] also studied learning from soft labels. However, the performance of these sequential models can be limited in scenarios requiring the modeling of complex, non-sequential dependencies between distant utterances and intricate speaker interactions [2,3]. This inherent challenge motivates the exploration of graph-structured representations.

2.1.2. Graph-Based Models

To explicitly model intra- and inter-speaker dependencies, graph neural networks have been widely adopted. Ghosal et al. [2] introduced DialogueGCN, constructing conversation graphs with distinct edge types. Joshi et al. [3] presented COGMEN, a hybrid architecture interleaving GNNs with graph-aware Transformers. Shen et al. [4] proposed DAG-ERC to model spatiotemporal dynamics via directed acyclic graphs, while Jia et al. [14] developed SIGAT, a unified heterogeneous graph fusing speaker and temporal information. Zhang et al. [15] introduced DualGATs to jointly exploit discourse structure and speaker context. For multimodal scenarios, Hu et al. [16] proposed MMGCN, and recently, Tu et al. [17] developed AdaIGN, which dynamically selects informative nodes and edges. A common characteristic of these graph-based methods is their reliance on predefined and relatively static graph architectures. While effective, this design may present a limitation in adapting to the unique and dynamically changing contextual demands of diverse conversational tasks.

2.1.3. Knowledge-Enhanced Models

Recognizing that human emotion is often grounded in common sense, another line of research incorporates external knowledge. Zhong et al. [18] introduced KET to augment representations with external common sense. Ghosal et al. [19] presented COSMIC, distilling speaker-specific common sense. Zhu et al. [20] devised TODKAT, fusing common sense with topic information, and Fu et al. [21] proposed CONSK-GCN for multimodal knowledge integration. Li et al. [22] introduced SKIER to model discourse relations with symbolic knowledge, Xie et al. [23] developed KI-Net with an auxiliary intensity prediction task, and Tu et al. [24] presented MKE-IGN for multi-knowledge enhancement. These methods demonstrate the benefits of incorporating external knowledge. However, this integration also introduces new considerations, such as computational overhead, potential noise, and the challenge of seamless fusion with the internal conversational context.

2.2. Contrastive Learning for ERC

Contrastive learning has recently gained traction for learning discriminative utterance representations by leveraging the relational nature of emotions [25,26]. Gunel et al. [27] proposed Supervised Contrastive Learning (SupCon), and Kang et al. [28] introduced KCL to address class imbalance. Li et al. [29] devised TSC to guide compact clustering, while Zhu et al. [30] and Zhang et al. [31] incorporated learnable prototypes and label embeddings into the contrastive loss, respectively. In conversational scenarios, Tu et al. [32] presented CKCL, a context-knowledge contrastive framework, and Yu et al. [33] introduced emotion-anchored contrastive learning (EACL). Notably, the contrastive objectives in these methods are often based on fixed heuristics for constructing positive and negative pairs. This fixed strategy struggles to capture the nuanced and task-specific relational structures in diverse ERC scenarios, highlighting the need for more adaptive contrastive mechanisms.

In summary, prior work has established robust paradigms for contextual modeling in ERC. A common observation, however, is the reliance on fixed architectural designs for capturing context, regardless of the specific conversation task. This reveals the need for a framework capable of synergistically and adaptively modeling both the macroscopic narrative flow and microscopic interaction patterns in conversations, and integrating this adaptability into its contextual modeling and learning objective.

2.3. Task-Adaptive Methods for ERC

Task-adaptive methods have emerged as a significant strategy in emotion recognition in conversations (ERC), with the goal of bridging the semantic gap between general-purpose pre-trained language models (PLMs) and the specific demands of the ERC task. These methods integrate task-specific knowledge (e.g., emotion category systems and conversational dynamics) during model training. By doing so, they guide PLMs to learn representations that are more discriminative for emotion recognition.

Existing task-adaptive approaches for ERC can be broadly categorized into three main strands. First, prompt-based methods, exemplified by Sentiprompt [34] and the framework of Cai et al. [35], utilize carefully designed, emotion-aware templates to reformat the input. This directs the model's focus toward emotion-relevant aspects of the dialogue. However, the static nature of such prompts may limit their ability to fully model dynamic emotional transitions across conversation turns. Second, contrastive learning frameworks enhance adaptation by introducing contrastive objectives during representation learning. EACL [33], for instance, constructs emotion-anchored vectors to cluster representations of the same emotion class. A potential limitation is that its task adaptation is somewhat decoupled from the downstream classification objective. SCCL [36] employs a three-dimensional (speaker, content, and sequential) representation to construct contrastive pairs. Yet, its heuristic-based sample construction may not fully capture the rich semantic relationships within the emotion space. Although SupCon-ERC [27] effectively applies a supervised contrastive loss to ERC, its design lacks an explicit, built-in mechanism for task adaptation, which could impact generalization across diverse conversational settings. Third, knowledge-enhanced methods, such as TODKAT [20] and KI-Net [23], incorporate external common sense or topical knowledge to enrich the dialogue's semantic context. This integration, however, typically increases computational complexity and introduces the challenge of seamlessly combining external knowledge with the intrinsic conversational context.

In summary, a key limitation of existing methods lies in their compartmentalized adaptation mechanisms, which are often not coherently integrated throughout the learning pipeline. Furthermore, the structured semantics of emotion categories are rarely systematically utilized to inform the model's learning objective. These gaps highlight the need for a unified and synergistic adaptive framework that can achieve coherent task adaptation

while explicitly leveraging the structured emotion space. Addressing these gaps is therefore key to advancing towards more theoretically grounded and empirically robust solutions in ERC.

3. Problem Definition

We formalize a conversation as comprising an utterance sequence and a corresponding speaker sequence:

$$\begin{aligned} U &= \{u_1, u_2, \dots, u_n\} \\ S &= \{s_1, s_2, \dots, s_n\} \end{aligned} \quad (1)$$

where each u_i is uttered by speaker s_i . Here, the conversation involves m distinct speakers ($m \leq n$), with each s_i representing one unique speaker identity.

The emotion recognition in conversations (ERC) task requires predicting the emotion label L_e^i for every utterance $u_i \in U$. Each label L_e^i is drawn from the dataset's emotion space E , which defines all admissible emotion categories—for instance, $E = \{\text{neutral, happy, angry, sad, frustrated, excited}\}$ in the IEMOCAP benchmark.

For ease of explanation, relevant notations used in this article are summarized in Table 1.

Table 1. Mathematical notations.

Symbol	Description
U	A conversation, composed of multiple utterances.
S	The set of speakers in a conversation.
E	The emotion space of the dataset, for instance, $E = \{\text{neutral, happy, angry, sad, frustrated, excited}\}$ in the IEMOCAP benchmark.
e_i	The emotion concept vector corresponding to an emotion category in E_r .
E_r	The set of all emotion concept vectors, $\{e_1, e_2, \dots, e_n\}$.
p	A prompt template.
MLP_{ce}	The projection head for contrastive learning.
d_r	The representation of an utterance.
Z_r	The projected utterance representation for contrastive learning.
τ	Temperature hyperparameter.
$\text{sim}(z_i, z_j)$	Cosine similarity function.
W_t	A local context window centered on the t -th utterance.
C_t^{self}	Historical utterances of the target speaker (up to time t).
C_t^{inter}	Adjacent utterance pairs within the local context.

4. The TAMC-ERC Model

To tackle the core challenges in Emotion Recognition in Conversations defined in Section 3 (namely, the semantic gap between pre-trained language models and the task-specific emotion space, and the effective integration of global coherence and local interactions in conversations), this paper presents task-adaptive and multi-level contextual modeling (TAMC-ERC). The proposed framework is designed to achieve precise transformation from general semantic understanding to task-specific emotion discrimination through an end-to-end, progressive learning architecture. Specifically, the framework first injects emotion concepts into utterance representations via task-adaptive representation learning. It then constructs a comprehensive emotional context that synergistically integrates macroscopic narratives and microscopic interactions. Finally, a classifier parameterized by the emotion concept vectors transforms the refined representations into emotion predictions. The following sections detail the design rationale and technical specifications of each core component.

4.1. Model Framework Overview

To systematically address the progressive research problem identified in Section 1, this paper presents the TAMC-ERC model (illustrated in Figure 1). The model follows a “utterance–context–classification” architecture: it first learns foundational representations at the utterance level, then advances to contextual understanding at the conversation level, and finally achieves task-adaptive classification. Through this architecture, TAMC-ERC systematically tackles the two challenges outlined in the Introduction via three core components.

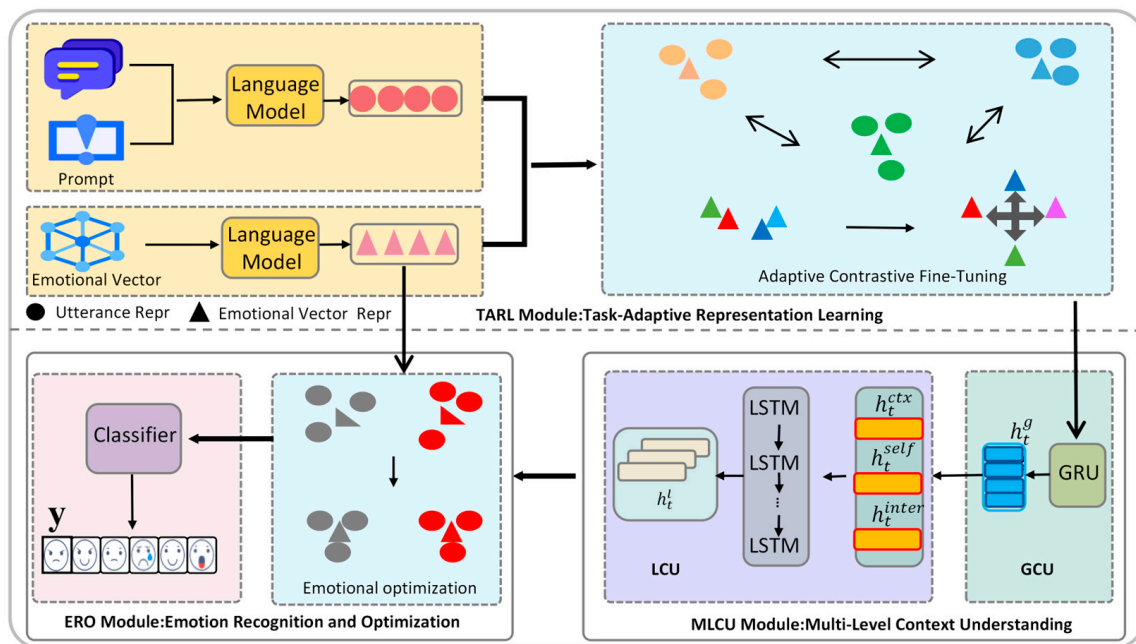


Figure 1. The overall architecture of the TAMC-ERC model.

First, the Task-Adaptive Representation Learning module (Section 4.2) addresses the alignment between the PLM and the emotion space. Operating at the utterance level, this module leverages emotion space-constrained prompt encoding and contrastive fine-tuning to generate a high-quality foundational representation for each utterance that is highly discriminative within the emotion space. The output of this stage serves as a critical semantic foundation for the entire system and is fed into subsequent modules.

Subsequently, the Multi-Level Contextual Understanding module (Section 4.3) tackles the integration of global coherence and local interactions in conversations. Operating at the conversation level, this module processes information from two perspectives: the Global Contextual Understanding captures the macroscopic narrative structure and long-range dependencies via a bidirectional GRU, while the Local Contextual Understanding captures fine-grained interactive dynamics by modeling the local semantic field, speaker state trajectories, and direct interaction adjacency. The fusion of their outputs provides rich contextual information for each utterance, resulting in a contextually refined representation.

Finally, the Emotion Recognition and Optimization module (Section 4.4) performs the classification. A key design of this module is to share the emotion concept vectors with the task-adaptive stage and use them to parameterize the classifier. This ensures the task adaptation from representation learning to the final decision. The entire model is trained end-to-end by combining classification and contrastive learning objectives.

In conclusion, the TAMC-ERC model combines three core components: task-adaptive representation learning, multi-level contextual understanding, and a classifier parameter-

ized by emotion concept vectors, thereby establishing a progressive reasoning paradigm for conversational emotion recognition.

4.2. Task-Adaptive Representation Learning

This module addresses the core challenge of adapting general-purpose pre-trained language models (PLMs) to the emotion recognition in conversations (ERC) task, where PLMs' lack of intrinsic knowledge of the structured emotion space E hinders their discriminative capability. The proposed solution is a task-adaptive learning module that operates at the utterance level, explicitly leveraging the emotion space E as a semantic constraint to specialize PLM representations for ERC. The module operates in two cohesive stages: (1) task-adaptive prompt context encoding (Section 4.2.1) achieves an initial, static adaptation by incorporating E directly into the input; (2) task-adaptive contrastive fine-tuning (Section 4.2.2) builds upon this foundation to address the persistent ambiguity between semantically close emotion categories in the initially adapted representations, dynamically refining them using E .

4.2.1. Task-Adaptive Prompt Context Encoding

Standard prompt-based methods often fall short in ERC due to their unrestricted output space, which is ill-suited for predicting a fixed set of emotions. To bridge this gap, this paper proposes a task-adaptive prompt that explicitly integrates the task dataset's emotion space E as a semantic constraint, effectively closing the PLM's output space.

Following the standard offline ERC benchmark, we construct the prompt using only the left context of the target utterance. This design is supported by empirical findings across prompt-based ERC studies, which show that emotional cues are predominantly triggered by preceding utterances, whereas subsequent utterances may contain sarcasm or emotional flip, acting as label noise (Li et al. [34], 2021; Cai et al. [35], 2025 findings). Accordingly, for a target utterance u_t , its historical context (comprising all preceding utterances and speakers) is utilized. This modeling approach aligns with conversational dynamics and ensures model practicality for real-time applications. Based on this historical context, the input sequence x_t is constructed by appending a structured prompt P :

$$x_t = [u_1, s_1, \dots, u_t, s_t; P] \quad (2)$$

Here, $u_1, s_1, \dots, u_t, s_t$ denote the sequence of utterances and speakers in the historical context, and P is the task-adaptive prompt. P is formulated as "Emotions in this conversation are confined to the set E . For utterance u_t , speaker s_t feels the emotion [mask]". In our experiments, E is instantiated with the emotion label set of the specific benchmark dataset, e.g., {happy, sad, angry, ...} for IEMOCAP. This design provides the PLM with explicit knowledge of valid emotion categories, establishing task-adaptive semantic constraints that significantly reduce classification ambiguity for conversational emotion recognition.

The initial task-aware representation for the target utterance is derived from the hidden state of the [mask] token:

$$\mathcal{Z}_{[mask]} = PLM(x_t)_{[mask]} \quad (3)$$

Here, $\mathcal{Z}_{[mask]} \in \mathbb{R}^d$ denotes the contextualized embedding for a single target utterance, serving as the initial task-aware representation for subsequent ERC steps. For a batch of b utterances, we obtain $\mathcal{Z}_{[mask]} \in \mathbb{R}^{b \times d}$ where b is the batch size, and d is the hidden size of the PLM (i.e., 1024 for Sup-SimCSE-RoBERTa-large).

4.2.2. Task-Adaptive Contrastive Fine-Tuning

Building on the preliminary utterance-level representations from Section 4.2.1, this stage aims to enhance their discriminative power. Although the task-adaptive prompt encoding provides preliminary task awareness, its static nature results in representations that are often insufficient for fine-grained emotional distinctions (e.g., between ‘anger’ and ‘frustration’). To overcome this, we introduce a task-adaptive contrastive fine-tuning module. This component directly leverages the emotion space E of the task dataset through contrastive learning. The objective is to transform the initial task-aware representations into ones that are highly discriminative within the target emotion space.

To project the representations into a space optimized for similarity comparison, the representation $Z_{[mask]}$ is transformed into a lower-dimensional embedding space using a multi-layer perceptron (MLP):

$$Z_r = \text{MLP}_{ce}(Z_{[mask]}) \quad (4)$$

where MLP_{ce} denotes the projection layer for contrastive learning, and $Z_r \in \mathbb{R}^{b \times d_r}$ represents the refined utterance representations for the batch, with d_r being the dimension of the contrastive embedding space.

To ground the contrastive-learning process, we first generate a concept vector for each emotion category in the task-specific set E . This vector is obtained by an independent MLP projection layer (MLP_r) that operates on the output of the PLM. Importantly, the PLM remains fully frozen throughout training:

$$E_r = \text{MLP}_r(\text{PLM}(E)) \quad (5)$$

where $E_r \in \mathbb{R}^{|E| \times d_r}$ represents the emotion concept vectors, and $|E|$ is the number of emotion categories. Subsequently, to construct a unified set for contrastive learning, we combine the utterance representations Z_r with the emotion concept set E_r . Formally, this combined set is constructed as follows:

$$Z_a = \text{torch.cat}([Z_r, E_r], \text{dim} = 0) \quad (6)$$

$$Y_a = \text{torch.cat}([Y, Y_e], \text{dim} = 0) \quad (7)$$

Here, $Z_a \in \mathbb{R}^{(b + |E|) \times d_r}$ denotes the concatenated representation matrix, where b is the batch size, $|E|$ is the number of emotion categories, and a is equal to $b + |E|$. $Y \in \mathbb{R}^b$ contains the ground-truth emotion labels for the b utterances in the batch, and $Y_e \in \mathbb{R}^{|E|}$ contains the predefined labels for each vector in the emotion concept set E_r .

The supervised contrastive loss is then computed on this unified set Z_a . The pairwise similarity between two representations z_i and z_j in Z_a is calculated as

$$c_{ij} = \frac{\text{sim}(z_i, z_j)}{\tau} \quad (8)$$

where $\text{sim}(z_i, z_j)$ denotes the cosine similarity function, and τ is a temperature hyperparameter that controls the concentration level of the distribution.

The supervised contrastive loss $\mathcal{L}_{NCE}$ on the unified set Z_a is formulated as

$$\mathcal{L}_{NCE} = -\frac{1}{|Z_i^+|} \sum_{i=1}^a \log \frac{\sum_{z_j \in Z_i^+} e^{c_{ij}}}{\sum_{z_j \in Z_a} e^{c_{ij}}} \quad (9)$$

where $Z_i^+ = \{z_j \in \mathcal{Z}_a \mid y_j = y_i, j \neq i\}$ denotes the set of positive samples (i.e., those sharing the same emotion label), for instance i , while the denominator sums over all samples in the unified set $\mathcal{Z}_a$ except i itself. This loss enhances the model's discriminative power by attracting representations of the same emotion class closer in the latent space, while repelling those from different classes.

To preserve the model's direct classification capability and provide a complementary supervisory signal to contrastive learning, a standard cross-entropy classification loss is jointly optimized. The emotion prediction distribution and the loss are formulated as follows:

$$\hat{y} = \text{softmax}(\text{MLP}_{\text{cs}}(\mathcal{Z}_r)) \quad (10)$$

$$\mathcal{L}_{CE} = -\frac{1}{b} \sum_{i=1}^b \sum_{j=1}^{|E|} \mathcal{Y}_{ij} \log \hat{\mathcal{Y}}_{ij} \quad (11)$$

where MLP_{cs} is the classification head, $\hat{y} \in \mathbb{R}^{b \times |E|}$ is the predicted probability distribution over the $|E|$ emotion classes for the batch, and $\mathcal{Y}_{ij}$ is the ground-truth label (1 if sample i belongs to class j and 0 otherwise).

The final optimization objective combines both learning signals to achieve a balance between discriminative feature learning and accurate classification. To further sharpen the decision boundaries within the emotion space E , a separation loss $\mathcal{L}_E$ is applied to the emotion concept vectors, which minimizes the maximum pairwise cosine similarity among them. This loss promotes a more dispersed distribution of vectors in the feature space by minimizing the maximum pairwise cosine similarity. Thus, the overall task-adaptive loss is defined as the weighted sum:

$$\mathcal{L}_E = \max_{\substack{i, j \in \{1, 2, \dots, |E|\} \\ i \neq j}} \text{sim}(e_i, e_j) \quad (12)$$

$$\mathcal{L}_{TA} = \lambda_C \mathcal{L}_{CE} + \lambda_N \mathcal{L}_{NCE} + \lambda_E \mathcal{L}_E \quad (13)$$

where $\lambda_C, \lambda_N, \lambda_E \in [0, 1]$ are the loss weighting coefficients that control the relative importance of the classification, contrastive learning, and emotion space separation objectives, satisfying $\lambda_C + \lambda_N + \lambda_E = 1$.

In summary, this component performs task-adaptive refinement of the representations through fine-tuning. Building upon the prompt encoding, it generates final utterance representations $\mathcal{Z}_r$ that are highly discriminative within the emotion space E of the task dataset. These refined representations are both context-aware and serve as a robust input for the subsequent ERC module.

4.3. Multi-Level Contextual Understanding

Accurate recognition of conversational emotion inherently depends on integrating both the global conversation flow and local inter-utterance interactions. To this end, this module employs a cascaded architecture designed to model these dual dynamics concurrently. The global module first encodes the entire conversation sequence to model the macro-level emotional flow. Subsequently, guided by this global context, the local module analyzes fine-grained interactions between adjacent utterances. This progressive, from-macro-to-micro comprehension ensures that the emotion of each utterance is derived from both its overall narrative position and its immediate conversational dynamics.

4.3.1. Global Contextual Understanding Module

Accurate emotion recognition in conversations requires understanding each utterance within its global conversational context, as emotions often evolve through long-range cross-

utterance dependencies. To this end, this module leverages the strength of gated recurrent units (GRUs) in sequence modeling to capture such dependencies. It feeds the utterance representation sequence $X = [x_1, x_2, \dots, x_n]$ from the first stage into a bidirectional GRU:

$$\begin{aligned} \vec{h}_t &= \vec{\text{GRU}}(x_t, \vec{h}_{t-1}) \\ \overleftarrow{h}_t &= \overleftarrow{\text{GRU}}(x_t, \overleftarrow{h}_{t+1}) \\ h_t^g &= [\vec{h}_t; \overleftarrow{h}_t], t = 1, 2, \dots, n \end{aligned} \quad (14)$$

Here, h_t^g denotes the global-contextualized representation of the t -th utterance, which encodes the macroscopic narrative structure of the conversation. This module outputs a sequence of global-contextualized representations $H^g = [h_1^g, h_2^g, \dots, h_n^g]$. This sequence serves as the input to the Local Contextual Understanding module, providing a representation foundation enriched with global context for subsequent fine-grained analysis. Unlike the left-context prompt that generates an initial emotion-aware representation for each utterance, the bidirectional GRU refines these representations by exploiting long-range dependencies across the entire conversation.

4.3.2. Local Contextual Understanding Module

Building upon the macroscopic context from the global module, this module aims to capture the diverse factors driving local emotional dynamics. Its design is grounded in the principle that utterance emotion is systematically influenced by the local “topic atmosphere,” the speaker’s “own state,” and the immediate “conversation partner”. Accordingly, the module employs a hierarchical modeling strategy, constructing local contexts of varying scopes to quantify these three key factors, respectively.

We begin by constructing a local context window of radius r around the global representation h_t^g of the t -th utterance. This window, extracted from the global representation sequence H^g , forms the basic unit for analysis.

$$W_t^g = [h_{t-r}^g, \dots, h_t^g, \dots, h_{t+r}^g] \quad (15)$$

Here, this window W_t^g contains the global-contextualized representations of the utterances most relevant to the current one, providing the foundational data for local context understanding.

(1) Foundational Context: Local Semantic Field

The local window W_t^g provides the complete semantic environment nearest to the current utterance and establishes the local topical background for emotional expression.

(2) Focused Context: Speaker State Trajectory

To capture a speaker’s emotional state continuity, we define a binary speaker mask M^{self} . Its elements M_{ij}^{self} are set to 1 if speaker s_j is the target speaker s_t and 0 otherwise. Applying this mask to the local window W_t^g isolates the speaker’s state trajectory C_t^{self} , as formalized in Equation (16).

$$M_{ij}^{\text{self}} = \begin{cases} 1 & \text{if } s_j = s_t \\ 0 & \text{otherwise} \end{cases}, \quad C_t^{\text{self}} = M^{\text{self}} \odot W_t^g \quad (16)$$

where s_j and s_t are the speaker identities of the j -th and t -th utterances, respectively. This operation yields a sequence C_t^{self} containing only the target speaker’s utterances, aimed at modeling individual emotional inertia.

(3) Core Context: Direct Interaction Adjacency Pair

To capture the most direct emotional interaction in conversation, we apply an adjacency mask to extract utterance pairs that constitute direct responses to the current utterance within a local window:

$$M_{ij}^{\text{inter}} = \begin{cases} 1 & \text{if } u_j \text{ is direct response to } u_t \\ 0 & \text{otherwise} \end{cases}, \quad C_t^{\text{inter}} = M^{\text{inter}} \odot W_t^g \quad (17)$$

This operation yields the crucial adjacency pair C_t^{inter} , which is designed to model the most direct emotional stimulus and response relationship.

To integrate the contextual features, the module processes three distinct sequences with separate LSTMs: the local window W_t^g , the speaker-state sequence C_t^{self} , and the adjacency-pair sequence C_t^{inter} . This is formalized in Equation (18):

$$h_t^{\text{ctx}} = \text{LSTM}(W_t^g), \quad h_t^{\text{self}} = \text{LSTM}(C_t^{\text{self}}), \quad h_t^{\text{inter}} = \text{LSTM}(C_t^{\text{inter}}) \quad (18)$$

By processing the full local context, speaker-specific trajectory, and adjacency pairs through separate LSTMs, the module extracts their respective temporal features h_t^{ctx} , h_t^{self} , and h_t^{inter} .

Finally, the outputs of the LSTMs are concatenated to form the final local fusion representation for the i -th utterance:

$$h_t^l = [h_t^{\text{ctx}}; h_t^{\text{self}}; h_t^{\text{inter}}] \in \mathbb{R}^{3d} \quad (19)$$

where d denotes the dimensionality of a single utterance representation. This fused representation h_t^l systematically integrates the three types of key contextual information: the local semantic field, the speaker state trajectory, and the direct interaction adjacency pair. The resulting representation thereby forms a fine-grained, multifaceted local context that provides crucial micro-level dynamic evidence, complementing the macro-level cues from the global context for final emotion determination.

4.4. Emotion Recognition and Optimization

This module performs the final emotion classification prediction through the systematic integration of the outputs of the preceding modules. It leverages both the context-enriched utterance representations from the multi-level context understanding module (Section 4.3) and the emotion concept vectors, which capture the core semantics of each emotion category, derived from the task-adaptive contrastive learning (Section 4.2).

To achieve direct alignment with the task-specific semantic space, we construct the final classifier by directly parameterizing it with the emotion concept vectors E_r . For a context-rich utterance representation h_i , this classifier assigns the label of the most similar concept vector $e_j \in E_r$. This design explicitly grounds the classification in the semantic space established during task adaptation. Formally, for a given utterance representation h_i , the classification is performed by computing its similarity to all emotion concept vectors, as follows:

$$\hat{y}_i = \arg \max_{j \in \{1, \dots, C\}} \sigma(h_i, e_j) \quad (20)$$

Here, $e_j \in E_r$ denotes the emotion concept vector for the j -th emotion category, and $\sigma(h_i, e_j)$ is the cosine similarity function. This design ensures that the classification logic stems directly from the task-adaptive emotion concepts, rather than from traditional, data-agnostic model parameters.

The entire model is optimized end-to-end by minimizing the cross-entropy loss between the predictions and the ground-truth emotion labels:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C 1_{[y_i=j]} \log(p_i^j) \quad (21)$$

where N is the batch size, $1_{[y_i=j]}$ is the indicator function, and p_i^j is the probability distribution obtained by applying the Softmax function to the similarity scores.

5. Experiments

5.1. Datasets

In this work, we evaluate our model on three publicly available benchmarks for emotion recognition in conversations (ERC). The details of the datasets are as follows:

- (1) IEMOCAP Dataset [37]: This is a multimodal dataset. Each dialogue segment is performed by two actors based on scripts. Only the text modality is used in this experiment. The data samples in this dataset involve 6 emotion categories: happy, sad, angry, frustrated, excited, and neutral. Considering that there is no validation set in the IEMOCAP dataset, following the practice in DAG-ERC, we split the data into a training set and a validation set, with the last 20 dialogues held out for validation.
- (2) MELD Dataset [38]: This is a multi-party, multi-modal dialogue dataset derived from the TV show *Friends*. Different from IEMOCAP, this dataset encompasses seven emotion categories: anger, disgust, fear, happiness, neutral, sadness, and surprise.
- (3) EmoryNLP Dataset [39]: Similar to MELD, EmoryNLP is based on *Friends*, involving multi-party dialogues and containing 7 categories. However, EmoryNLP differs from MELD in scene selection, and the set of 7 emotion categories it defines differs from that of MELD, namely happy, powerful, peaceful, sad, crazy, fear, and neutral.

The statistical analysis of the above datasets is shown in Table 2. The variations in conversation structure and emotion categories across these datasets facilitate a comprehensive evaluation of model generalizability.

Table 2. Key statistics of the benchmark conversation datasets for ERC.

Dataset	Total Number of Conversations	Number of Emotion Categories	Average Number of Speakers per Conversation	Average Number of Clauses per Conversation
IEMOCAP	151	6	2	49.2
MELD	1433	7	9	9.6
EmoryNLP	897	7	9	14.1

5.2. Experimental Setup

Hyperparameters were searched on the validation set of each dataset (Table 3). Additionally, all experiments were conducted on a single RTX 4090 GPU manufactured by NVIDIA Corporation, based in Santa Clara, CA, USA, and the reported results are the average of five random runs on the test set.

Table 3. Hyperparameters of TAMC-ERC for three benchmark datasets.

Hyperparameters	IEMOCAP	MELD	EmoryNLP
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Dropout	0.2	0.2	0.2
Temperature	0.2	0.3	0.2

Table 3. *Cont.*

Hyperparameters	IEMOCAP	MELD	EmoryNLP
Maximum length	256	256	256
Batch size	32	32	32
Epochs	8	6	6

5.3. Baseline Models

For a comprehensive evaluation, our approach is compared against the following baseline models:

(1) Graph-based Models

DialogueGCN (Ghosal et al., 2019) [2] adopts GCN to capture contextual features for learning utterance representations. RGAT (Ishiwatari et al., 2020) [40] introduces relational position encoding to model speaker relationships and sequential information. DAG-ERC (Shen et al., 2021) [4] exploits an acyclic graph neural network to intuitively characterize the inherent structure of conversations without incorporating external information. DAG-ERC+HCL (Yang et al., 2022) [10] strengthens DAG-ERC by integrating a curriculum learning strategy, which trains the model on samples arranged from easy to difficult. SIGAT (Jia et al., 2023) [14] encodes speaker and sequence information into a unified graph to capture their mutual interactive effects. AdaIGN (Tu et al., 2024) [17] proposes an adaptive interaction graph network (IGN), which employs the Gumbel Softmax technique to adaptively select nodes and edges, thereby enhancing intra-modal and cross-modal interactions.

(2) Sequence-based Models

Cog-BART (Li et al., 2022) [5] utilizes BART combined with contrastive learning for the joint tasks of emotion recognition and response generation. DialogueEIN (Liu et al., 2022) [8] includes emotion interaction and tendency modules to explicitly model emotional inertia and stimuli. CoMPM (Lee and Lee, 2021) [41] leverages pre-trained models to directly acquire contextual information and track conversation history. Sup-Con (Gunel et al., 2020) [27] is a supervised contrastive learning loss function. SCCL (Yang et al., 2023) [36] conducts contrastive learning using three-dimensional influence representations. DIEU (Zhao et al., 2023) [42] aims to resolve the issue of long-distance context propagation. CKCL (Tu et al., 2023) [32] filters out context- and knowledge-irrelevant noise during the training process. MPLP (Zhang et al., 2023) [43] models speakers' historical experiences and employs paraphrasing to amplify the differences between labels. Emocaps (Li et al., 2022) [44] redesigns the Transformer into a novel architecture called the Bistable Transformer to extract the emotional tendency of utterances. SACL (Hu et al., 2023) [7] learns robust representations by fusing contrastive learning with adversarial training. SPCL+CL (Song et al., 2022) [45] integrates prototype contrastive learning and curriculum learning to address the problem of imbalanced emotion classes. ChatGPT-3.5 (Zhao et al., 2023) [46], whose 3-shot performance is reported as a reference. EACL (Yu et al., 2024) [33] is proposed using an emotion-anchored contrastive learning framework for discriminative utterance representations for similar emotions.

(3) Knowledge-Enhanced Models

KET (Zhong et al., 2019) [18] is a knowledge-enhanced Transformer framework that dynamically utilizes external common sense knowledge and attention mechanisms to model and understand dialogue emotions. COSMIC (Ghosal et al., 2020) [19] incorporates various commonsense elements and leverages them to learn self-speaker dependencies. TODKAT (Zhu et al., 2021) [20] integrates commonsense knowledge with

dialogue topic information to enhance adaptability across multiple scenarios. KI-Net (Xie et al., 2021) [23] combines multi-task learning for emotion polarity intensity prediction to strengthen knowledge interaction.

6. Experimental Results and Analysis

To comprehensively validate the design principles of the TAMC-ERC model, this section presents evaluations through three systematic experiments: We first establish its overall performance advantage by comparing with baseline models, then analyze the contribution of each core component through ablation studies, and finally, we investigate the synergistic working mechanism among key modules.

6.1. Overall Performance Comparison

Table 4 presents the weighted average F1 scores of TAMC-ERC and baseline models on three benchmark datasets.

Table 4. Performance comparison between TAMC-ERC and various baseline models.

Methods	IEMOCAP	MELD	EmoryNLP	Average
Graph-based models				
DialogueGCN (Ghosal et al., 2019) [2]	64.91	63.02	38.10	55.34
RGAT (Ishiwatari et al., 2020) [40]	66.36	62.80	37.89	55.68
DAG-ERC (Shen et al., 2021) [4]	68.03	63.65	39.02	56.90
DAG-ERC+HCL (Yang et al., 2022) [10]	68.73	63.89	39.82	57.48
SIGAT (Jia et al., 2023) [14]	70.71	66.20	39.95	58.77
AdaIGN (Tu et al., 2024) [17]	70.74	66.79	-	-
Sequence-based models				
CKCL (Tu et al., 2023) [32]	67.16	66.21	40.23	57.87
Cog-BART (Li et al., 2022) [5]	66.18	64.81	39.04	56.68
DialogueEIN (Liu et al., 2022) [8]	68.93	65.37	38.92	57.74
CoMPM (Lee and Lee., 2021) [41]	69.46	66.52	38.93	58.30
SupCon (Gunel et al., 2020) [27]	68.14	65.63	39.28	57.68
Emocaps (Li et al., 2022) [44]	69.49	63.51	-	-
SPCL+CL (Song et al., 2022) [45]	67.19	65.74	39.52	57.48
SACL (Hu et al., 2023) [7]	69.22	66.45	39.65	58.44
SCCL (Yang et al., 2023) [36]	69.88	65.70	38.75	58.11
DIEU (Zhao et al., 2023a) [42]	69.90	66.43	40.12	58.11
MPLP (Zhang et al., 2023) [43]	66.65	66.51	-	-
ChatGPT 3-shot (Zhao et al., 2023) [46]	48.58	58.35	35.92	47.62
EACL (Yu et al., 2024) [33]	70.41	67.12	40.24	59.26
Knowledge-Enhanced Models				
KET (Zhong et al., 2019) [18]	59.56	58.18	33.95	50.56
COSMIC (Ghosal et al., 2020) [19]	65.25	65.21	38.11	56.19
TODKAT (Zhu et al., 2021) [20]	61.33	65.47	43.12	56.53
KI-Net (Xie et al., 2021) [23]	66.98	63.24	-	-
TAMC-ERC (ours)	71.04	66.95	40.99	59.66

The experimental results demonstrate that TAMC-ERC achieves competitive performance across all datasets, obtaining the best results on IEMOCAP and performing comparably to advanced baselines on MELD and EmoryNLP. This robust performance highlights the efficacy of our architecture in handling conversational complexity and enabling multi-level contextual understanding. This is achieved through two key designs: (1) the Task-Adaptive Representation Learning module, which explicitly aligns the model with the task-specific semantic space via emotion concept vectors; and (2) the cascaded Global and Local Contextual Understanding modules, which systematically capture long-

range dependencies and fine-grained interactions. The synergy between these components collectively contributes to accurate and robust emotion recognition.

The observed performance variations across datasets align with their inherent complexity levels. On IEMOCAP, a dyadic dataset with well-defined speaker roles, TAMC-ERC demonstrates clear advantages. This result confirms the effectiveness of the intra-speaker modeling mechanism in the Local Contextual Understanding (LCU) module for such scenarios. Although all models exhibit performance degradation on the more challenging multi-speaker datasets (MELD and EmoryNLP), TAMC-ERC maintains robust performance, comparable to state-of-the-art approaches like SIGAT [14] and EACL [33].

To statistically validate the comparison with the strongest baseline, paired *t*-tests were conducted between TAMC-ERC and EACL. The improvement across all three datasets was not statistically significant ($p = 0.149$). However, a focused test on IEMOCAP and EmoryNLP, where TAMC-ERC showed superior performance, yielded a significant result ($p = 0.0037$). This indicates that while EACL is a highly competitive benchmark, our model's integrated architecture delivers robust and significant gains where its design is most effective.

Notably, the knowledge-enhanced model TODKAT [20] showed strong performance on EmoryNLP, likely benefiting from external commonsense knowledge integration. In contrast, TAMC-ERC achieved competitive results without external knowledge. This demonstrates that architectural innovations in task adaptation and multi-level contextual modeling are, by themselves, sufficient to substantially advance conversational emotion recognition. Furthermore, the results were consistent across multiple independent training runs, indicating robustness to random initialization. Regarding computational efficiency, TAMC-ERC introduced minimal overhead compared to standard PLM fine-tuning. The additional components—the contrastive loss $\mathcal{L}_{NCE}$, the emotion separation loss $\mathcal{L}_E$, and the lightweight context encoders—contributed to a negligible parameter increase (<2% of the PLM's size). Consequently, the inference speed remained practically unchanged, ensuring the model's feasibility for real-world applications.

6.2. Ablation Experiments

Ablation studies were conducted to validate the contribution of each core component in the TAMC-ERC framework. The experimental design corresponds to the two main innovations: the task-adaptive mechanism and multi-level contextual understanding. Table 5 shows the weighted average F1 scores on three benchmark datasets, where “↓” indicates performance degradation.

Table 5. Ablation experimental results of the TAMC-ERC model on each dataset.

Dataset	IEMOCAP	MELD	EmoryNLP
Ours	71.04	66.95	40.99
<i>w/o</i> Task-Adaptive Contrastive Learning	68.77 (2.27 ↓)	66.2 (0.75 ↓)	40.7 (0.29 ↓)
<i>w/o</i> Prototype-Parameterized	69.45 (1.59 ↓)	63.24 (3.71 ↓)	36.61 (4.38 ↓)
<i>w/o</i> Global Contextual Understanding	70.84 (0.20 ↓)	66.79 (0.16 ↓)	40.81 (0.18 ↓)
<i>w/o</i> Local Contextual Understanding	70.8 (0.24 ↓)	66.82 (0.13 ↓)	40.77 (0.22 ↓)
<i>w/o</i> Global-Local Model	69.64 (1.4 ↓)	66.26 (0.69 ↓)	38.92 (2.07 ↓)

↓ indicates the amount of decrease.

We first evaluated the task-adaptive mechanism. Disabling the contrastive learning component led to performance degradation of 2.27% on IEMOCAP and 0.75% on MELD. This decline underscores the module's critical role in acquiring discriminative semantic representations. Furthermore, removing the emotion concept vectors from the classification layer caused more substantial performance drops (3.71% on MELD and 4.38% on

EmoryNLP). Critically, this ablation severs the direct link between the adapted concept vectors and the decision layer. The consequent performance drop confirms that using E_r as classifier parameters contributes to robustness. This advantage is likely achieved by preserving decision boundaries aligned with the contrastively structured embedding space. These findings collectively demonstrate that the task-adaptive integration of emotion concepts across both representation learning and classification is a cornerstone of the model's effectiveness.

The ablation of either the global or local contextual modules led to modest but consistent performance degradation (averaging 0.18% and 0.20%, respectively). To thoroughly investigate their synergistic contribution and address potential concerns regarding the statistical significance of these improvements, we conducted an additional ablation where the global–local integration mechanism is removed. This ablation results in substantially larger performance drops: 1.40% on IEMOCAP, 0.69% on MELD, and 2.07% on EmoryNLP. Paired *t*-tests between the full model and this ablated variant confirm statistical significance (one-tailed $p = 0.037$, $t = 3.475$, $df = 2$), indicating over 95% confidence in the improvement. This stark contrast reveals that while individual modules have limited impact due to functional complementarity, their integrated modeling is essential for holistic conversation comprehension. The global module captures narrative-level flow and long-range dependencies, while the local module refines turn-level interaction dynamics. Their synergy provides the foundational context required by the task-adaptive mechanism, validating the multi-level contextual architecture as a core component of our framework.

6.3. Analysis of Synergistic Effects

To validate the synergistic effects between the global and local contextual modules in TAMC-ERC, we conducted a controlled experiment. We replaced the standard global–local interaction mechanism with a local-only variant, which processes adjacent utterances via LSTM without the global GRU-based context modeling. A performance comparison between these two architectures across three datasets is presented in Figure 2.

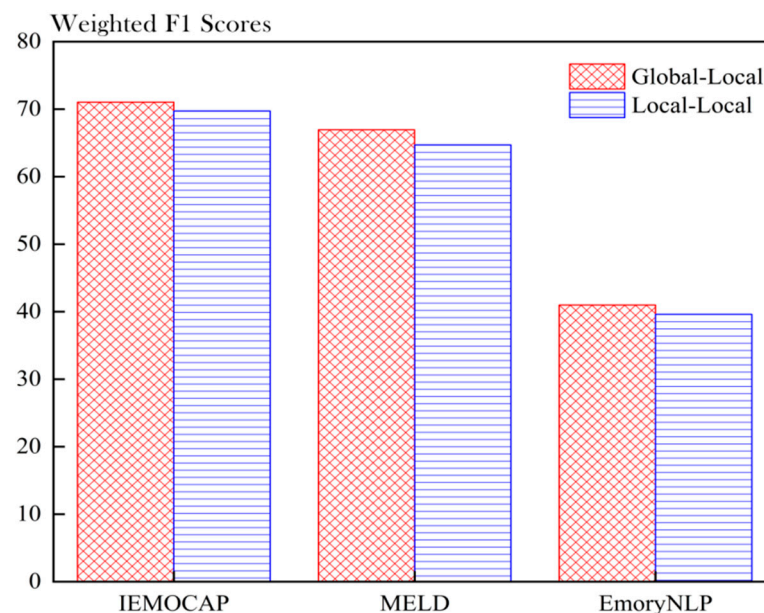


Figure 2. Performance advantage of the synergistic global–local model.

Experimental results indicate that the global–local interaction mechanism consistently outperforms the local-only baseline across all datasets. This performance gap occurs because modeling long-range emotional evolution across conversation turns is essential.

However, the local-only interactions prove insufficient for this task, as they primarily capture surface-level semantic associations between adjacent utterances. The synergistic effect is achieved precisely through the distinct and complementary roles of these two modules. The Local Contextual Understanding module captures fine-grained details like speaker identity and adjacent relationships, providing crucial cues for refining the emotion concept vectors. In parallel, the Global Contextual Understanding module encodes the overarching conversation semantics, thereby setting the directional constraints that ensure all local adjustments remain coherent with the primary emotional trajectory.

This collaborative mechanism proves particularly important in semantically complex multi-party conversations. For instance, in a multi-party debate scenario, local interactions capture immediate rebuttals among participants, while the global context frames the broader emotional progression, such as a shift from “debate → banter → reconciliation”. This enables the dynamic adjustment of emotion representations based on local cues while ensuring their consistency with the global emotional narrative.

7. Conclusions

This work addressed two pivotal challenges in conversational emotion recognition: bridging the semantic gap between general-purpose pre-trained models and specialized emotion tasks and synergistically modeling global and local contextual information. Our proposed TAMC-ERC model tackles these through a synergistic integration of task-adaptive and multi-level contextual modeling. The model follows a progressive cognitive pipeline. First, the Task-Adaptive Representation Learning module injects the structure of the emotion space via prompts and contrastive learning, yielding highly discriminative utterance representations. Subsequently, the Multi-level Contextual Understanding module systematically integrates macro-level conversation flow and micro-level interactive dynamics through cascaded global and local modeling. Finally, the classifier is parameterized by the learned emotion concept vectors, ensuring semantic alignment with the task-specific space. Extensive experiments confirm TAMC-ERC’s competitive advantage, with ablation studies underscoring the indispensable role and superadditive synergy of both core components. Together, these findings establish the deep integration of task adaptation and multi-level contextual modeling as an effective paradigm for conversational emotion recognition.

However, experimental analysis shows that TAMC-ERC still faces challenges in comprehending complex scenarios within multi-party conversations (e.g., the EmoryNLP dataset). We will therefore develop effective multimodal fusion strategies to address this limitation. Beyond improving the specific model, the task-adaptive paradigm established by TAMC-ERC shows potential for extension to other conversational AI scenarios facing challenges of task-specificity and complex context. For example, the framework’s core principles are transferable to other sequential or textual analysis tasks requiring similar capabilities. Promising applications include fault diagnosis in building systems [47], where the method could model temporal sequences by adapting to fault concepts and capturing hierarchical contextual patterns, as well as automated classification of construction inspection reports [48], where task-adaptive prompts could encode defect types, and contextual modules could interpret technical descriptions within their narrative structure.

Author Contributions: Conceptualization, W.C.; methodology, W.C. and X.Y.; software, W.C. and X.Y.; validation, X.Y. and Y.X.; formal analysis, X.Y. and Y.X.; investigation, X.Y. and Y.X.; writing—original draft preparation, X.Y.; writing—review and editing, W.C.; visualization, X.Y.; supervision, H.Z. and X.F.; funding acquisition, W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was mainly supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region, China (Grant No. 2022D01A228); the Doctoral Research Foundation of Xinjiang Normal University (Grant No. XJNUZBS2441); the Bidding Project of Xinjiang Engineering Research Center for Intelligent Education and Applications (Grant No. XJNU-ZHJY202407); and the Education Reform Project of Xinjiang Normal University (Grant No. SDJG2024-30).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study were derived from three public benchmark datasets for conversational emotion recognition: IEMOCAP, MELD, and EmoryNLP. All datasets are accessible in their respective public repositories.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. DialogueRNN: An attentive RNN for emotion detection in conversations. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 6818–6825.
2. Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; Gelbukh, A. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, 3–7 November 2019; pp. 154–164.
3. Joshi, A.; Bhat, A.; Jain, A.; Singh, A.V.; Modi, A. Cogmen: Contextualized gnn based multimodal emotion recognition. *arXiv* **2022**, arXiv:2205.02455. [CrossRef]
4. Shen, W.; Wu, S.; Yang, Y.; Quan, X. Directed acyclic graph network for conversational emotion recognition. *arXiv* **2021**, arXiv:2105.12907. [CrossRef]
5. Li, S.; Yan, H.; Qiu, X. Contrast and generation make bart a good dialogue emotion recognizer. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 11002–11010.
6. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7871–7880.
7. Hu, D.; Bao, Y.; Wei, L.; Zhou, W.; Hu, S. Supervised adversarial contrastive learning for emotion recognition in conversations. *arXiv* **2023**, arXiv:2306.01505.
8. Liu, Y.; Zhao, J.; Hu, J.; Li, R.; Jin, Q. Dialogueein: Emotion interaction network for dialogue affective analysis. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 684–693.
9. Yang, H.; Gao, X.; Wu, J.; Gan, T.; Ding, N.; Jiang, F.; Nie, L. Self-adaptive Context and Modal-interaction Modeling For Multimodal Emotion Recognition. In *Proceedings of the Findings of the Association for Computational Linguistics, ACL 2023, Toronto, ON, Canada, 9–14 July 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 6267–6281.
10. Yang, L.; Shen, Y.; Mao, Y.; Cai, L. Hybrid curriculum learning for emotion recognition in conversation. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022; Volume 36, pp. 11595–11603.
11. Lei, S.; Dong, G.; Wang, X.; Wang, K.; Wang, S. Instruclerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv* **2023**, arXiv:2309.11911.
12. Zhang, Y.; Wang, M.; Tiwari, P.; Li, Q.; Wang, B.; Qin, J. Dialogueellm: Context and emotion knowledge-tuned llama mod-els for emotion recognition in conversations. *arXiv* **2023**, arXiv:2310.11374.
13. Lee, J. The emotion is not one-hot en-coding: Learning with grayscale label for emo-tion recognition in conversation. *arXiv* **2022**, arXiv:2206.07359.
14. Jia, Z.; Shi, Y.; Liu, W.; Huang, Z.; Sun, X. Speaker-aware interactive graph attention network for emotion recognition in conversation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–18. [CrossRef]
15. Zhang, D.; Chen, F.; Chen, X. Dualgats: Dual graph attention networks for emotion recognition in conversations. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; Volume 1: Long Papers, pp. 7395–7408.
16. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in con-versation. *arXiv* **2021**, arXiv:2107.06779.

17. Tu, G.; Xie, T.; Liang, B.; Wang, H.; Xu, R. Adaptive graph learning for multimodal conversational emotion detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 26–27 February 2024; Volume 38, pp. 19089–19097.
18. Zhong, P.; Wang, D.; Miao, C. Knowledge-enriched transformer for emotion de-tection in textual conversations. *arXiv* **2019**, arXiv:1909.10681.
19. Ghosal, D.; Majumder, N.; Gel-bukh, A.; Mihalcea, R.; Poria, S. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv* **2020**, arXiv:2010.02795.
20. Zhu, L.; Pergola, G.; Gui, L.; Zhou, D.; He, Y. Topic-driven and knowledge-aware transformer for dialogue emotion detection. *arxiv* **2021**, arXiv:2106.01071.
21. Fu, Y.; Okada, S.; Wang, L.; Guo, L.; Dang, J. CONSK-GCN: Conversational semantic-and knowledge-oriented graph convolutional network for multimodal emotion recognition. In *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021*; IEEE Computer Society: Washington, DC, USA, 2021; pp. 1–6.
22. Li, W.; Zhu, L.; Mao, R.; Cambria, E. SKIER: A symbolic knowledge integrated model for conversational emotion recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Montréal, QC, Canada, 8–10 August 2023; Volume 37, pp. 13121–13129.
23. Xie, Y.; Yang, K.; Sun, C.; Liu, B.; Ji, Z. Knowledge-interactive network with emotion polarity intensity-aware multi-task learning for emotion recognition in conversations. In Proceedings of the Findings of the Association for Computational Linguistics, EMNLP 2021, Virtual, 16–20 November 2021; pp. 2879–2889.
24. Tu, G.; Wang, J.; Li, Z.; Chen, S.; Liang, B.; Zeng, X.; Yang, M.; Xu, R. Multiple knowledge-enhanced interactive graph network for multimodal conversational emotion recognition. In Proceedings of the Findings of the Association for Computational Linguistics, EMNLP 2024, Miami, FL, USA, 12–16 November 2024; pp. 3861–3874.
25. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning, Virtual, 13–18 July 2020*; PMLR: Breckenridge, CO, USA, 2020; pp. 1597–1607.
26. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
27. Gunel, B.; Du, J.; Conneau, A.; Stoyanov, V. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv* **2020**, arXiv:2011.01403.
28. Kang, B.; Li, Y.; Xie, S.; Yuan, Z.; Feng, J. Exploring balanced feature spaces for representation learning. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020.
29. Li, T.; Cao, P.; Yuan, Y.; Fan, L.; Yang, Y.; Feris, R.S.; Indyk, P.; Katabi, D. Targeted supervised contrastive learning for long-tailed recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6918–6928.
30. Zhu, J.; Wang, Z.; Chen, J.; Chen, Y.-P.P.; Jiang, Y.-G. Balanced contrastive learning for long-tailed visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6908–6917.
31. Zhang, Z.; Zhao, Y.; Chen, M.; He, X.-A. Label anchored contrastive learn-ingfor language understanding. *arXiv* **2022**, arXiv:2205.10227.
32. Tu, G.; Liang, B.; Mao, R.; Yang, M.; Xu, R. Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 14054–14067.
33. Yu, F.; Guo, J.; Wu, Z.; Dai, X. Emotion-anchored contrastive learning framework for emotion recognition in conversation. *arxiv* **2024**, arXiv:2403.20289.
34. Li, C.; Gao, F.; Bu, J.; Xu, L.; Chen, X.; Gu, Y.; Shao, Z.; Zheng, Q.; Zhang, N.; Wang, Y.; et al. Sentiprompt: Sentiment knowledge enhanced prompt-tuning for aspect-based sentiment analysis. *arXiv* **2021**, arXiv:2109.08306.
35. Cai, C.; Zhang, K.; Hu, Z.; Lin, X.; Pan, Z. Prompt-based hybrid supervised conetrative learning for emotion recognition in conversation. *Neurocomputing* **2025**, *647*, 130453.
36. Yang, K.; Zhang, T.; Alhuzali, H.; Ananiadou, S. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Trans. Affect. Comput.* **2023**, *14*, 3269–3280. [[CrossRef](#)]
37. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359.
38. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 527–536.

39. Zahiri, S.M.; Choi, J.D. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In Proceedings of the Workshops at the thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 44–51.
40. Ishiwatari, T.; Yasuda, Y.; Miyazaki, T.; Goto, J. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 7360–7370.
41. Lee, J.; Lee, W. Compm: Context modeling with speaker’s pre-trained memory track-ing for emotion recognition in conversation. *arXiv* **2021**, arXiv:2108.11626.
42. Zhao, S.; Liu, W.; Chen, J.; Sun, X. Dieu: A dynamic interaction emotion unit for emotion recognition in conversation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2023**, *22*, 1–18.
43. Zhang, T.; Chen, Z.; Zhong, M.; Qian, T. Mimicking the thinking process for emotion recognition in conversation with prompts and paraphrasing. *arXiv* **2023**, arXiv:2306.06601. [[CrossRef](#)]
44. Li, Z.; Tang, F.; Zhao, M.; Zhu, Y. Emocaps: Emotion capsule based model for conversational emotion recognition. *arXiv* **2022**, arXiv:2203.13504. [[CrossRef](#)]
45. Song, X.; Huang, L.; Xue, H.; Hu, S. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv* **2022**, arXiv:2210.08713. [[CrossRef](#)]
46. Zhao, W.; Zhao, Y.; Lu, X.; Wang, S.; Qin, B. Is chat-gpt equipped with emotional dialogue capabilities? *arXiv* **2023**, arXiv:2304.09582.
47. Wang, S. Real operational labeled data of air handling units from office, auditorium, and hospital buildings. *Sci. Data* **2025**, *12*, 1481. [[CrossRef](#)]
48. Wang, S.; Moon, S.; Eum, I.; Hwang, D.; Kim, J. A text dataset of fire door defects for pre-delivery inspections of apartments during the construction stage. *Data Brief* **2025**, *60*, 111536. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.