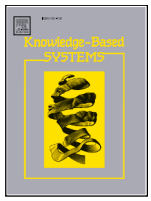
Knowledge-Based Systems 340 (2026) 115594

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Cross-modality multiband differential conditional diffusion for multimodal emotion recognition in conversation

Xiaofei Zhu [a,*], Yang Jiang [a], Xiaoyang Liu [a], Yihao Zhang [b]

[a] School of Computer Science and Engineering, Chongqing University of Technology, Chongqing, 400054, China
[b] School of Artificial Intelligence, Chongqing University of Technology, Chongqing, 400054, China

ARTICLE INFO

Keywords:
Multimodal emotion recognition
Diffusion model
Discrete wavelet transform
Emotion analysis

ABSTRACT

Multimodal Emotion Recognition in Conversation (MERC) aims to identify emotional category of each utterance by leveraging complementary cues from text, audio, and visual modalities. Despite recent advances, existing multimodal fusion methods often suffer from cross-modal interference and inadequate handling of utterance-specific noise. To address these limitations, we propose a novel Cross-Modality Multiband Differential Conditional Diffusion (CMDCD) framework. Specifically, our approach introduces an asymmetric fusion enhancement (AFE) mechanism to enhance representation learning of each modality. It alternately treats each modality as primary, using others as auxiliaries to provide complementary information and reduce cross-modal interference. Moreover, we develop a wavelet differential conditional diffusion (WDCD) module for multiband denoising. It first decomposes features into frequency subbands via wavelet transformation and then constructs differential representations to capture utterance-specific features, which are fed into a conditional diffusion process guided by cross-band signals to suppress utterance-specific noise while preserving fine-grained emotional cues. At last, we leverage a confidence-based fusion (CBF) strategy to further integrate the asymmetric fusion enhanced features and the multiband denoised features based on their prediction confidence. Extensive experiments on two widely utilized benchmarks demonstrate that CMDCD consistently outperforms the state-of-the-art methods. We have provided the source code of our proposed CMDCD model at the following link: https://github.com/madaler/CMDCD.

1. Introduction

Multimodal Emotion Recognition in Conversation (MERC) aims to identify emotions in utterances by analyzing the dialogue content, bringing significant benefits to intelligent human-computer interaction [1], course quality assessment [2], and social media opinion mining [3]. Early methods employ Recurrent Neural Networks (RNNs) and Graph Convolutional Networks (GCNs) to identify emotions within the textual modality [4–6]. Since emotions are conveyed across multiple modalities, including text, audio, and video, relying solely on text is inadequate for a comprehensive understanding of conversational emotion [1].

Recent methods leverage sequential and graph structures to integrate text, audio, and video modalities for enhancing emotion detection [7]. For instance, CTNet [8] employs unimodal and cross-modal transformer to capture both intra- and inter-modal dependencies. MPT-HCL [9] adopts a multimodal prompt transformer with hybrid contrastive learning to fuse filtered multimodal cues and handle low-resource emotion categories. AdaIGN [10] utilizes GCNs to fuse multimodal features and applies Gumbel-Softmax to adaptively select nodes and edges, thereby enhancing the effectiveness of modality interactions. HAUCL [11] introduces a hypergraph autoencoder to learn multimodal information and long-range contextual dependencies while mitigating redundancy and over-smoothing issues in conventional graph models. Besides, SEDC [12] introduces a dual-channel architecture to decouple the processing of semantic and emotional information. The model employs contrastive learning to distill emotional features from each utterance, while simultaneously leveraging an external knowledge base to enrich the semantic representation of the dialogue.

Although these methods achieve promising results, they still suffer from the following limitations: (1) *Cross-modal interference*. Existing multimodal fusion methods usually rely on a symmetric fusion strategy, where all modalities are treated equally [8,11]. This strategy often leads to mutual interference of information across modalities, thereby compromising the effectiveness of fusion [13]. (2) *Inadequate handling of utterance-specific noise*. Existing multimodal fusion approaches primarily focus on capturing inter-modal dependencies to enhance fusion performance, while paying less attention to the impact of noise within individual utterance during the fusion process. It may lead to the amplification

* Corresponding author.
E-mail addresses: zxf@cqut.edu.cn (X. Zhu), jiangy@stu.cqut.edu.cn (Y. Jiang), lxy3103@cqut.edu.cn (X. Liu), yhzhang@cqut.edu.cn (Y. Zhang).

https://doi.org/10.1016/j.knosys.2026.115594
Received 25 December 2025; Received in revised form 4 February 2026; Accepted 19 February 2026
Available online 26 February 2026

of negative effects during the fusion process, as noise from each utterance propagates to other modalities and ultimately degrades the final fusion results. Although recent research efforts, such as diffusion models, have achieved promising results in modality denoising, these methods mainly operate on original tangled features where salient information and noise are non-trivially intertwined [14]. Specifically, each utterance inherently comprises two types of components: *utterance-common features*, which represent the core emotional semantics shared across the same category, and *utterance-specific features*, which encapsulate the unique nuances and noise specific to an individual utterance. It is worth noting that these utterance-specific features usually serve as the primary source of noise. Existing methods perform denoising directly on original features without distinguishing between them, which limits denoising effectiveness.

To address the aforementioned issues, we propose a novel Cross-Modality Multiband Differential Conditional Diffusion (CMDCD) framework for multimodal emotion recognition in conversation. The framework is composed of three key modules: (1) *Asymmetric Fusion Enhancement (AFE)*. To alleviate the cross-modal interference caused by symmetric fusion mechanism, we propose an asymmetric fusion enhancement module that alternately treats each modality as the primary one, while the others act as auxiliaries for guiding integration. This design enables the model to learn more discriminative and expressive representations for the main modality while preserving complementary cues across modalities. (2) *Wavelet Differential Conditional Diffusion (WDCD)*. Unlike traditional methods, we propose to perform fine-grained multiband denoising. Specifically, we decompose the original feature space into frequency subbands via wavelet transformation, enabling more fine-grained handling of noise across different frequency domains. Besides, we introduce emotion semantic prototypes to represent utterance-common features and construct differential representations to characterize utterance-specific features, which are then fed into a conditional diffusion process to suppress the utterance-specific noise. To enhance the denoising process of each subband space, we further leverage signals from other subbands to guide the denoising procedure. (3) *Confidence-based Fusion (CBF)*. To harness the complementary strengths of above two modules, i.e., AFE and WDCD, we employ an adaptive fusion mechanism according to their prediction confidences, and combine the asymmetric fusion enhanced and multiband denoised features for optimal performance.

In summary, our contributions are as follows:

- We propose an asymmetric fusion enhancement mechanism to alleviate cross-modal interference, which alternately treats each modality as the primary source and leverages auxiliary modalities to enhance main modality representation.
- We propose a wavelet differential conditional diffusion model for fine-grained, multiband denoising. It decomposes the feature space into frequency subbands via wavelet transformation and selectively suppress noise in utterance-specific features by constructing differential representations and leveraging cross-band conditional diffusion guidance to supervise the denoising procedure.
- We employ a confidence-based fusion strategy to adaptively integrate asymmetric fusion enhanced and multiband denoised features, leveraging their complementary strengths to improve model effectiveness and robustness.
- Extensive experiments on two benchmark datasets demonstrate that our method outperforms state-of-the-art baselines in terms of both accuracy and weighted F1-score (Fig. 1).

2. Related work

2.1. Multimodal emotion recognition in conversation

Emotion recognition in conversation has garnered increasing attention due to its wide applications across various fields. Existing studies could be broadly categorized into sequence-based and graph-based approaches, both aiming to model contextual dependencies and multimodal interactions. Sequence-based methods primarily relied on recurrent or transformer architectures to capture temporal and contextual dependencies. DialogueRNN [4] first proposed an RNN-based approach that tracked the speaker's emotional state and contextual information through collaborative modeling of Global GRU, Party GRU, and Emotion GRU. CTNet [8] adopted a Transformer-based architecture combining multi-head attention and bidirectional GRU to capture both unimodal and cross-modal dependencies. MM-DFN [15] introduced a gating mechanism to dynamically aggregate contextual information across modalities, effectively reducing redundancy and enhancing complementarity. CMCF-SRNet [16] integrated text and audio modalities through cross-modal local constraint transformers and semantic graph transformers, improving contextual understanding and recognition accuracy. However, while sequence-based methods effectively capture temporal dependencies, modeling the complex and often non-sequential dependencies across modalities remains a challenge.

To more flexibly capture the structured interactions and multimodal relationships within conversations, many graph-based approaches have been proposed. DialogueGCN [5] employed a GCN to model the conversation graph, optimizing the propagation of contextual information. By using a two-layer GCN, it strengthened emotional correlations between speakers. MMGCN [17] constructed a multimodal graph convolutional network to facilitate the interaction of information from different modalities. M3Net [18] utilized hypergraph neural networks for multi-variable propagation, enhancing emotion recognition accuracy. HAUCL [11] proposed a hypergraph-based cross-modal emotion recognition framework. By constructing hypergraph convolutions and introducing contrastive learning, it effectively integrated multimodal information. AdaIGN [10] optimized the graph structure via an adaptive interaction graph network and employed a self-supervised method to generate pseudo-labels, thereby improving emotion recognition accuracy. LECM [19] leveraged emotion causes in real-time ERC by combining a causal emotion entailment task with a cause-oriented tendency network, improving recognition of subtle emotional cues. FrameERC [20] used a frame let transform-based multimodal GNN to capture low- and high-frequency emotional signals and enhanced non-textual modality contributions via a dual-reminder fusion mechanism. MATCH [21] proposed a modality-calibrated hypergraph fusion network that calibrated utterance and speaker representations and constructed an emotion-aligned hypergraph to capture cross-modal and cross-utterance dependencies. SEDC [12] proposed a semantic and emotional dual channel strategy, by leveraging contrastive learning and knowledge enhancement, the model effectively mitigated semantic noise and emotional confusion. Despite their encouraging progress, these methods usually struggle with cross-modal interference and insufficient noise modeling. To address these issues, we propose an asymmetric fusion enhancement mechanism and a wavelet differential conditional diffusion model to suppress cross-modal interference and utterance-specific noise.

2.2. Diffusion model

In recent years, diffusion models had garnered significant attention in the field of generative modeling [22,23], with widespread applications in areas such as recommendation systems [24] and multimodal tasks [25,26]. Conventional models primarily consisted of two paradigms: Denoising Diffusion Probabilistic Models (DDPM) [27] and Score-based Generative Models (SGM) [28]. Currently, numerous studies explored the application of diffusion models in emotion recognition tasks. For example, IMDer [29] introduced a score-based diffusion to address the issue of missing modalities in emotion recognition. TopicDiff [30] integrated a denoising diffusion model with neural topic model to alleviate the lack of semantic diversity in traditional topic models when capturing multimodal topic information. RMER-DT [31] proposed a conditional diffusion framework to iteratively reconstruct

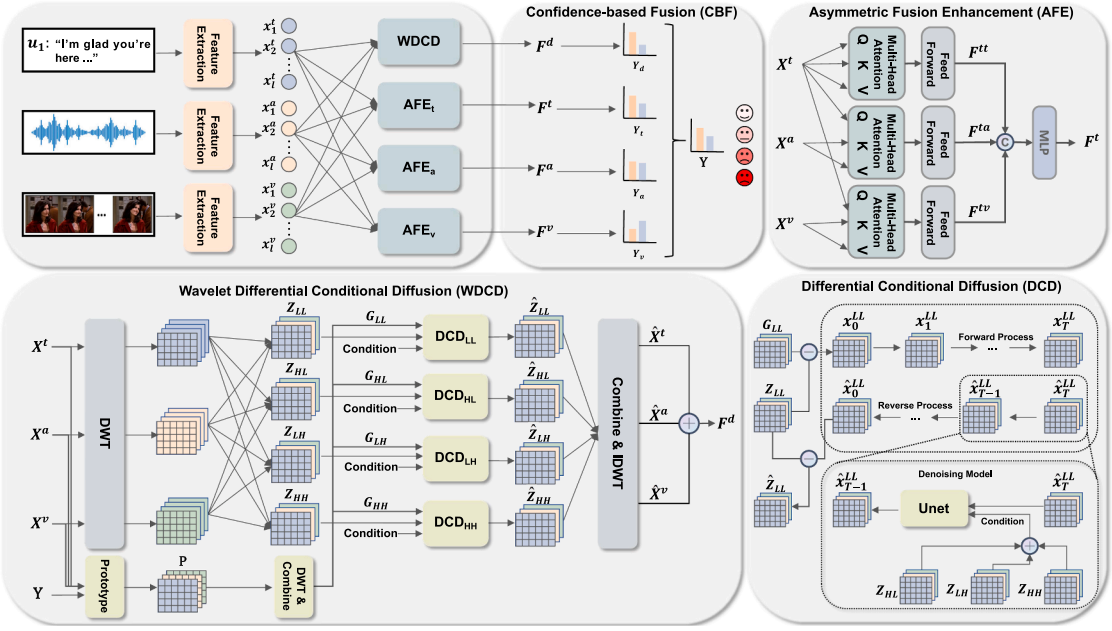


Fig. 1. Illustration of the CMDCD framework, which consists of three key components: (1) Asymmetric Fusion Enhancement (AFE), which alternately treats each modality as the primary source to reduce cross-modal interference and preserve modality-specific cues; (2) Wavelet Differential Conditional Diffusion (WDCD), which performs fine-grained multi-scale denoising by leveraging emotion semantic prototypes, differential representations, and cross-band conditional signals to suppress noise in utterance-specific features; and (3) Confidence-based Fusion (CBF), which adaptively integrates the enhanced features from AFE and WDCD, leveraging their complementary strengths to improve robustness and interpretability.

missing modalities and integrated the recovered features with a transformer-based fusion network, improving robustness under random modality missingness.

While recent diffusion-based methods have shown promising in handling missing modalities and generative tasks, their application in multimodal emotion recognition remains limited, where they mostly operate on the original tangled feature space rather than decomposed frequency spaces, and do not differentiate utterance-common and utterance-specific components. Our proposed model extends traditional diffusion frameworks by enabling fine-grained handling of noise across different frequency domains.

3. Methodology

3.1. Task definition and notation

Multimodal Emotion Recognition in Conversation (MERC) is defined as follows: given a conversation $U = \{u_1, u_2, \ldots, u_N\}$ containing N utterances, the model aims to predict the emotion label for each utterance in the conversation from a predefined set of emotion classes C. Each utterance consists of features from three modalities: text (t), audio (a) and vision (v), represented as $u_i = \{u_i^t, u_i^a, u_i^v\}, i \in \{1, \ldots, N\}$. Additionally, each utterance u_i is associated with its corresponding speaker s_{u_i}.

The overall framework of CMDCD is illustrated in Fig. 1. To facilitate understanding of the proposed CMDCD framework presented in the subsequent sections, Table 1 provides a summary of the key mathematical symbols used throughout the paper.

3.2. Feature extraction

The features of the text, audio, and visual modalities are first extracted from their respective raw data using the RoBERTa large model [32], the OpenSmile toolkit [33], and the DenseNet model [34], respectively. Then, a one-dimensional convolutional neural network (Conv1D) is used to transform the feature dimensions of the three modalities into a d-dimensional space. The speaker index is mapped to the embedding space, resulting in the speaker embedding S. Subsequently, the position embedding PE is incorporated, and the three are added as follows:

$$X^m = H^m + S + PE, m \in \{t, a, v\}, \tag{1}$$

where $H^m \in \mathbb{R}^{N \times d}$ denotes the modality feature after dimension alignment, the feature sequence X^m is used as the input for subsequent modules.

3.3. Asymmetric fusion enhancement (AFE)

In multimodal conversations, the quality of different modalities usually varies considerably, and conventional strategy of symmetric fusion would lead to undesired interference between modalities, resulting in suboptimal outcomes. To mitigate this issue, we design an asymmetric fusion enhancement module that alternately treats each modality as the primary one, while leveraging the remaining modalities as auxiliaries to provide complementary information. It allows the model to learn more discriminative representations for the main modality while preserving informative cues from other modalities.

Taking text as the primary modality, for example, we use the text features X^t as the query Q, with keys K and values V provided by X^t, X^a, and X^v, respectively. The process of integrating audio information into text features is formulated as:

$$F^{ta} = \mathrm{FFN}(\mathrm{Softmax}(X_t W_t^Q (X_a W_a^K)^T / \sqrt{d_h}) X_a W_a^V), \tag{2}$$

where W_t^Q, W_a^K, W_a^V are trainable projection matrices for Q, K, and V. FFN denotes a feed-forward network, and F^{ta} represents text features enhanced with audio cues. Similarly, we obtain text features fused with visual information F^{tv} and contextual features F^{tt}. These features are concatenated and projected to produce the final fusion feature F^t for text as the primary modality:

$$F^t = \mathrm{MLP}(\mathrm{Concat}(F^{tt}, F^{ta}, F^{tv})). \tag{3}$$

The same procedure is also applied to derive the audio-primary features F^a and the visual-primary features F^v.

Table 1
Summary of key notations.

Notation	Description
X^m	Conversation sequence feature of modality m (text, visual, or audio), where $m \in \{t, a, v\}$.
S	Speaker embedding obtained via speaker identity indexing.
PE	Positional encoding used to model the temporal order of utterances in a conversation.
Z_j	Multimodal frequency-domain conversational feature in subband $j \in \{LL, HL, LH, HH\}$ obtained via wavelet transform.
P^m	Emotion semantic prototype corresponding to modality m.
G_j	Multimodal frequency-domain conversational prototype feature corresponding to subband j.
$q(x_t \mid x_{t-1})$	The conditional distribution of x_t under the x_{t-1} condition in the forward process.
x_0	Initial input of the forward diffusion process.
x_t	Feature at step t in the forward diffusion process.
x_c	Conditional feature used to guide noise recovery.
β_t	Variance control parameter of Gaussian noise in the forward diffusion, with $\beta_t \in (0, 1)$.
$\alpha_t = 1 - \beta_t$	Proportion of the original signal retained at timestep t, with $\alpha_t \in (0, 1)$.
$\bar{\alpha}_t$	Cumulative product of α_t up to timestep t.
ϵ_t	The standard Gaussian noise added in the t-th step of forward diffusion.
$p_\theta(x_{t-1} \mid x_t, x_c)$	The conditional distribution of x_{t-1} under the conditions of x_t and x_c predicted in the reverse diffusion process.
μ_θ	The conditional mean of the reverse diffusion distribution x_{t-1} predicted by the neural network.
σ_t^2	The fixed variance of the distribution of x_{t-1} in the reverse process.
ϵ_θ	The estimated noise predicted by the neural network at timestep t.
a_i^k	Contribution weight of utterance i in modality k.

3.4. Wavelet differential conditional diffusion (WDCD)

Multimodal features often contain noise that varies across different spectral spaces, which can degrade emotion recognition performance if not handled properly. To address this issue, we propose a novel wavelet differential conditional diffusion module that performs multiband denoising in the frequency domain based on the discrete wavelet transform technique. By decomposing features into multiple frequency subbands, our model can selectively suppress utterance-specific noise.

3.4.1. Discrete wavelet decomposition

The discrete wavelet transform (DWT) is widely used in signal processing and low-level vision tasks [35,36]. In particular, we apply the Haar wavelet [37], which decomposes the signal using two orthogonal filters: low-pass filter L and high-pass filter H. The former smooths the signal, while the latter detects local changes or discontinuities. The filters are defined as follows:

$$L = \frac{1}{\sqrt{2}}[1, 1]^T, H = \frac{1}{\sqrt{2}}[1, -1]^T. \tag{4}$$

Given the feature representation $X^m \in \mathbb{R}^{N \times d}$ for the three modalities, we apply DWT to decompose the features of each modality into four frequency subbands:

$$[X^m_{LL}, X^m_{HL}, X^m_{LH}, X^m_{HH}] = \mathrm{DWT}(X^m), \tag{5}$$

where X^m_{LL} reflects low-frequency information, while X^m_{HL}, X^m_{LH}, and X^m_{HH} correspond to high-frequency information in the horizontal, vertical, and diagonal directions, respectively. These subbands are downsampled to half the input dimension without any loss of information due to the biorthogonal property of DWT [38]. To model the frequency-domain structure of multimodal information more effectively, we combine the corresponding frequency subbands from different modalities to form four multimodal frequency component features:

$$Z_j = \mathrm{Concat}(X^t_j, X^a_j, X^v_j), j \in \{LL, HL, LH, HH\}, \tag{6}$$

where $Z_j \in \mathbb{R}^{3 \times \frac{N}{2} \times \frac{d}{2}}$ is used to construct the differential input for the subsequent diffusion enhancement module, enabling the model to learn both contextual and multimodal features in the wavelet space.

3.4.2. Differential conditional diffusion

Existing denoising methods typically operate directly on original feature representations, and the denoising effect will be compromised due to both utterance-common and utterance-specific components are highly entangled. To alleviate this issue, we introduce differential representations to explicitly capture utterance-specific features. Specifically, based on the observation that utterances sharing the same emotion category exhibit consistent emotional semantics, we construct emotion semantic prototypes, which reflect utterance-common features, by averaging utterance features within each emotion class. Differential representations are then obtained by subtracting the corresponding emotion prototype from each utterance original feature representation. By applying diffusion-based denoising to these differential representations, the model can focus on suppressing utterance-specific noise, resulting in more effective denoising.

Specifically, we first compute the emotion semantic prototype P corresponding to all labels. Taking the text modality as an example, given the text modality features $X^t \in \mathbb{R}^{N \times d}$ and the corresponding emotion category labels $Y \in \mathbb{R}^N$, where $Y_i \in C$ denotes the emotion category corresponding to the i-th utterance in the emotion category set C. We compute average pooling on the features of all utterances with the same label in X^t to obtain the emotion semantic prototype for the conversational text modality P^t. For each emotion category, the corresponding emotion semantic prototype is computed as:

$$P^t_c = \frac{1}{|S_c|} \sum_{i \in S_c} X^t_i, \tag{7}$$

where $S_c = \{i \mid Y_i = c\}$ represents the index set of utterances with emotion category c in the conversation, and $|S_c|$ indicates the number of utterances with emotion category c. Similarly, P^a and P^v can be obtained. Following the procedure in Section 3.4.1, we perform DWT and frequency subband reorganization on the conversational emotion semantic prototype to obtain the corresponding frequency emotion semantic prototype:

$$[P^m_{LL}, P^m_{HL}, P^m_{LH}, P^m_{HH}] = \mathrm{DWT}(P^m), \tag{8}$$

$$G_j = \mathrm{Concat}(P^t_j, P^a_j, P^v_j), j \in \{LL, HL, LH, HH\}. \tag{9}$$

Taking the low-frequency subband LL as an example, the differential diffusion module can be divided into forward diffusion process and reverse diffusion process:

- **Forward Diffusion Process:** The forward diffusion process can be viewed as a Markov chain that gradually adds gaussian noise to the data. In this process, we construct a differential distribution $x^{LL}_0 = Z_{LL} - G_{LL}$ as the input. At each time step t, Gaussian noise is gradually added to the distribution as follows:

$$q(x^{LL}_t \mid x^{LL}_{t-1}) = \mathcal{N}(x^{LL}_t; \sqrt{1 - \beta_t} x^{LL}_{t-1}, \beta_t I), \tag{10}$$

where β_t controls the variance of the noise. Let $\alpha_t = 1 - \beta_t$, then the process is expressed as:

$$x_t^{LL} = \sqrt{\alpha_t} x_{t-1}^{LL} + \sqrt{1-\alpha_t}\epsilon_{t-1}, \epsilon_{t-1} \sim \mathcal{N}(0, \mathcal{Z}). \tag{11}$$

After merging the gaussian distribution, we obtain:

$$q(x_t^{LL}|x_0^{LL}) = \mathcal{N}(x_t^{LL}; \sqrt{\bar{\alpha}_t} x_0^{LL}, (1-\bar{\alpha}_t)I). \tag{12}$$

- **Reverse Diffusion Process:** The reverse diffusion process aims to recover the differential distribution from gaussian noise, which is expressed as:

$$p_\theta(x_{t-1}^{LL}|x_t^{LL}, x_c^{LL}) = \mathcal{N}(x_{t-1}^{LL}; \mu_\theta(x_t^{LL}, x_c^{LL}, t), \sigma_t^2 \mathcal{Z}), \tag{13}$$

where condition $x_c^{LL} = Z_{HL} + Z_{LH} + Z_{HH}$ is leveraged to guide the recovery of the noise added during the forward diffusion process. Previous studies have suggested that different frequency subbands provide complementary cues [38], with low-frequency components encoding global semantics and high-frequency components capturing fine-grained details. Cross-band conditioning exploits this complementarity to guide more effective denoising. $\mu_\theta(x_t^{LL}, x_c^{LL}, t)$ and σ_t^2 are the mean and variance estimated at step t, respectively, which are expressed as:

$$\mu_\theta(x_t^{LL}, x_c^{LL}, t) = \frac{1}{\sqrt{\alpha_t}}(x_t^{LL} - \frac{\beta_t}{(1-\bar{\alpha}_t)}\epsilon_\theta(x_t^{LL}, x_c^{LL}, t)), \tag{14}$$

$$\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \tag{15}$$

where $\epsilon_\theta(x_t^{LL}, x_c^{LL}, t)$ is the noise value estimated by the U-Net model. Specifically, the time-step information is first encoded by a positional embedding function and added to the x_t^{LL}, and then concatenated with the conditional feature along the channel dimension as the input to the U-Net:

$$\epsilon_\theta(x_t^{LL}, x_c^{LL}, t) = \text{U-Net}(\text{Concat}(x_t^{LL} + \text{PE}(t), x_c^{LL})). \tag{16}$$

To optimize the model, we minimize the error between the estimated noise and the noise added during the forward diffusion process. The loss function for the diffusion process is given by:

$$\mathcal{L}_{dm} = \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0^{LL} + \sqrt{1-\bar{\alpha}_t}\epsilon, x_c^{LL}, t)\|. \tag{17}$$

In the inference process, we use gaussian noise as the starting point for the reverse diffusion process. The denoised low-frequency features are obtained by subtracting the differential distribution $\hat{x}_0^{LL}$ generated by the diffusion model from the initial low-frequency features Z_{LL}:

$$\hat{Z}_{LL} = Z_{LL} - \hat{x}_0^{LL}. \tag{18}$$

Similarly, we can obtain the denoised features of the remaining frequencies. Subsequently, we split and reorganize the enhanced frequency-domain features and perform the inverse discrete wavelet transform (IDWT) to restore the features for the three modalities:

$$\hat{X}^m = \text{IDWT}(\hat{X}_{LL}^m, \hat{X}_{HL}^m, \hat{X}_{LH}^m, \hat{X}_{HH}^m), \tag{19}$$

where $\hat{X}^m$ is the result after multiband utterance-specific feature denoising. The final conversation feature is obtained by summing the denoised features from the three modalities:

$$F^d = \hat{X}^t + \hat{X}^a + \hat{X}^v. \tag{20}$$

3.5. Confidence-based fusion (CBF)

To effectively integrate the asymmetric fusion enhanced features and the multiband denoised features, we introduce a confidence-based fusion module. This strategy aims to adaptively balance the contributions of different proceeded representations according to their prediction confidence. Specifically, each enhanced or denoised representations independently predicts the emotion distribution through a corresponding classifier:

$$\hat{y}_i^k = \text{Softmax}(\text{MLP}_\mathbf{k}(F_i^k)), k \in \{d, t, a, v\}, \tag{21}$$

Algorithm 1 CMDCD framework training.

Input: Multimodal features $\{X^t, X^a, X^v\}$; Training labels Y; Max steps T

Output: Final predicted labels $\hat{Y}$

1. **Data Preparation**
   For each $m \in \{t, a, v\}$ **do**:
   (a) $P_c^m \leftarrow \text{MeanPooling}(X^m, Y), \forall c \in C$ // Generate prototypes via Eq. (7)
   (b) $\{X_j^m, P_j^m\}_{j\in\{LL,HL,LH,HH\}} \leftarrow \text{DWT}(X^m, P^m)$ // Multi-band decomposition via Eq. (5), Eq. (8)
   (c) $Z_j \leftarrow \text{Concat}(F_j^t, F_j^a, F_j^v)$, $G_j \leftarrow \text{Concat}(P_j^t, P_j^a, P_j^v)$
2. **Differential Conditional Diffusion(DCD)**
   For each $j \in \{LL, HL, LH, HH\}$ **do**:
   (a) $x_0^j \leftarrow G_j - Z_j$
   (b) **compute** $q(x_t^j|x_0^j)$ // Forward diffusion via Eq. (10)-(12)
   (c) $x_c^j \leftarrow \sum_{k\neq j} Z_k$ // Construct condition
   (d) $\hat{x}_T^j \sim \mathcal{N}(0, \mathbf{I})$ // Initialize Gaussian noise
   (e) **For** $t = T$ **down to** 1 **do**:
      i. $\epsilon_\theta \leftarrow \text{U-Net}(\hat{x}_t^j, x_c^j, t)$ // Estimate noise via Eq. (16)
      ii. $p_\theta(\hat{x}_{t-1}^j|\hat{x}_t^j, x_c^j) \leftarrow \text{ReverseStep}(\hat{x}_t^j, \epsilon_\theta, t)$ // compute va Eq. (13)-(15)
      iii. $\hat{x}_{t-1}^j \sim p_\theta(\hat{x}_{t-1}^j|\hat{x}_t^j, x_c^j)$ // Sample to obtain the next latent state
      iv. $\mathcal{L}_{dm} \leftarrow \mathcal{L}_{dm} + \text{MAE}(\epsilon_t, \epsilon_\theta)$
   (f) $\hat{Z}_j \leftarrow \hat{x}_0^j + Z_j$ // Restore denoised subband
3. **Wavelet Differential Conditional Diffusion**
   For each $m \in \{t, a, v\}$ **do**:
   (a) $\hat{Z}_j^m \leftarrow \text{DCD}(G_j^m, Z_j^m, condition)$
   (b) $\hat{X}_j^m \leftarrow \hat{Z}_j^m$ // split and reorganize
   (c) $\hat{X}^m \leftarrow \text{IDWT}(\hat{X}_{LL}^m, \hat{X}_{HL}^m, \hat{X}_{LH}^m, \hat{X}_{HH}^m)$
   (d) $F^d \leftarrow \hat{X}^t + \hat{X}^a + \hat{X}^v$.
4. **Asymmetric Fusion Enhancemen**
   (a) $F^m \leftarrow \text{AFE}(X^t, X^a, X^v), \forall m \in \{t, a, v\}$ // Eq. (2)-(3)
5. **Confidence-based Fusion**
   (a) $\hat{Y}, \mathcal{L} \leftarrow \text{CBF}(\{F^d, F^t, F^a, F^v\})$ // Eq. (22)-(24)
   (b) **Optimize the loss** $\mathcal{L}$
6. **Return** $\hat{Y}$

where $\hat{y}_i^k$ indicates the emotional prediction probability of the i-th utterance in the k-th classifier. The corresponding cross-entropy loss is:

$$\mathcal{L}_i^k = -\sum_j^C y_{ij} \log(\hat{y}_{ij}^k), \tag{22}$$

where y_i is the true label and j represents the j-th category. To better capture the contribution of each proceeded representation, we utilize a

Table 2
Statistics of IEMOCAP and MELD datasets.

Dataset	Dialogues			Utterances			Classes
	train	valid	test	train	valid	test	
IEMOCAP	120		31	5,810		1623	6
MELD	1039	114	280	9989	1109	2610	7

weight factor a_i^k for the i-th utterance in the k-th classifier:

$$a_i^k = \left\{ \prod_{j \neq k} \left(1 - \mathrm{MAX}(\hat{y}_i^j) \right) \right\}^{\delta}, \tag{23}$$

where the hyperparameter δ controls the learning degree of the weight factor, determining the utilization of multimodal data. Additionally, we use the maximum predicted probability $\mathrm{MAX}(\hat{y}_i^j)$ only during inference, while leverage the predicted probability corresponding to the ground-truth label during training. If a classifier shows better predictive performance, the weight of the other classifiers will be compressed, and the weight of the dominant classifier will be relatively amplified. Finally, the total prediction probability is the weighted sum of the individual classifier prediction probabilities.

Our proposed CMDCD framework is outlined in the Algorithm 1. The overall model loss is defined as:

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{k} a_i^k \mathcal{L}_i^k + \mathcal{L}_{dm}, k \in \{d, t, a, v\}. \tag{24}$$

4. Experiments

4.1. Dataset

To evaluate the effectiveness of the proposed approach CMDCD, we conduct experiments on two widely used MERC datasets, including **IEMOCAP** [39] and **MELD** [40]. The dataset statistics are summarized in Table 2.

- IEMOCAP: This dataset contains video recordings of 10 actors (5 male, 5 female) performing dyadic conversations. It includes 151 dialogues and 7433 utterances. The emotional labels are manually classified into six emotions: happy, sad, neutral, angry, excited, and frustrated.
- MELD: This is a multimodal dialogue dataset derived from the TV series Friends, containing 1433 dialogues and 13,708 utterances. Each utterance is labeled with one of seven emotion categories: neutral, surprise, fear, sadness, joy, disgust, and angry.

4.2. Evaluation metrics

Following previous studies [11,12], we utilize weighted F1-scores (WF1) and Accuracy (Acc.) as evaluation metrics. Note that the weighted F1 metric is employed to address the impact of class imbalance among different emotions, providing a fairer evaluation across all categories. In addition, we report per-class WF1 scores on the IEMOCAP dataset to offer a detailed analysis of model performance on each emotion category.

4.3. Experimental settings

All experiments are conducted using PyTorch 1.13.1 with CUDA 11.6 on a NVIDIA GeForce RTX 4090 GPU. and the Adam optimizer with L_2 weight decay is used to prevent overfitting. Each model is trained and evaluated five times with different random initializations, and the averaged results on the test set are reported. Most hyperparameters differ between IEMOCAP and MELD to account for dataset-specific characteristics. MELD contains longer dialogues and more complex multimodal interactions, so parameters such as attention heads, hidden dimensions, diffusion steps, learning rate, weight factor δ, and L2 regularization are adjusted accordingly to ensure effective modeling and stable optimization. The specific hyperparameter settings are shown in Table 3.

4.4. Baseline models

We conduct a comprehensive comparison of our proposed method with twelve state-of-the-art baseline methods, which can be grouped into two categories:

Sequence-based methods:

- DialogueRNN [4] models speaker states and contextual dependencies using a hierarchical RNN structure composed of Global, Party, and Emotion GRUs.
- CTNet [8] combines multi-head attention and bidirectional GRU within a Transformer-based framework to jointly capture unimodal and cross-modal dependencies.
- MM-DFN [15] introduces a dynamic gating mechanism to adaptively fuse contextual information across modalities, enhancing complementarity and reducing redundancy.
- CMCF-SRNet [16] employs cross-modal constraint transformers and semantic graph transformers to integrate text and audio, improving contextual representation.

Graph-based methods:

- DialogueGCN [5] constructs a conversation graph using GCNs to model speaker interactions and enhance emotional correlation propagation.
- MMGCN [17] constructs a multimodal graph convolutional network to facilitate interaction and information exchange among different modalities.
- M3Net [18] utilizes hypergraph neural networks for multi-variable propagation, thereby strengthening multimodal relationship modeling.
- HAUCL [11] designs a hypergraph-based cross-modal framework enhanced with contrastive learning to integrate multimodal features more effectively.
- AdaIGN [10] employs an adaptive interactive graph to enhance intra- and cross-modal interactions by selecting key nodes and edges, using directed graphs to prevent future utterances from influencing the current one.
- LECM [19] leverages emotion causes in real-time ERC by combining a causal emotion entailment task with a shared encoder and a cause-oriented tendency network, improving recognition of subtle emotional cues.
- FrameERC [20] decomposes graph signals into low- and high-frequency components for capturing fine-grained emotional cues, and employs a dual-reminder fusion mechanism to mitigate excessive dependence on the text modality.
- SEDC [12] adopts a dual-channel strategy to separately model semantic and emotional cues, leveraging contrastive learning and knowledge enhancement to mitigate semantic noise and emotional confusion.

4.5. Performance comparison

Table 4 presents the overall performance of our proposed method CMDCD and twelve baselines on two datasets (i.e., IEMOCAP and MELD). The best and second-best results in each column are highlighted in bold and underlined, respectively. From Table 4, we can observe that graph-based methods generally perform better than traditional sequence-based models due to their capability to capture cross-modal

Table 3
Main hyperparameters for CMDCD.

Parameters	IEMOCAP	MELD	Description
$batch\_size$	4	4	Number of conversation samples used in each training batch.
$epoch$	80	100	Total number of training epochs.
lr	1×10^{-4}	1×10^{-5}	Step size for parameter updates during optimization.
$l2$	7×10^{-5}	9×10^{-5}	L_2 regularization coefficient in Adam optimizer.
dim	320	400	Dimensionality of hidden feature representations.
$head\_num$	2	4	Number of attention heads used for modeling contextual interactions.
T	700	900	Number of timesteps in the diffusion denoising process.
δ	0.5	0.3	A scaling factor that controls the sensitivity of confidence-based fusion weights to prediction confidence.
β	linear	linear	Strategy for controlling noise variance across diffusion timesteps.

Table 4
Performance comparison of different methods on the IEMOCAP and MELD datasets (Bold values denote the best performance, and underlined values indicate the second-best).

Method	IEMOCAP								MELD	
	Emotion Categories(F1)						Overall		Overall	
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc.	WF1	Acc.	WF1
DialogueRNN	33.18	78.80	59.21	65.28	71.86	58.91	63.40	62.75	60.31	57.66
CTNet	51.30	79.90	65.80	67.20	<u>78.70</u>	58.80	–	67.00	–	60.50
MM-DFN	42.22	78.98	66.42	<u>69.77</u>	75.56	66.33	68.21	68.18	62.49	59.46
CMCF-SRNet	52.20	80.90	68.80	**70.30**	76.70	61.60	–	69.60	–	62.30
DialogueGCN	47.10	80.88	58.71	66.08	70.97	61.21	65.54	65.04	58.62	56.36
MMGCN	45.45	77.53	61.99	66.67	72.04	64.12	66.56	68.71	59.31	57.82
M3NET	57.96	81.56	68.30	65.59	74.91	63.19	69.01	69.12	67.62	66.15
HAUCL	53.57	<u>82.04</u>	68.61	66.44	75.60	68.23	70.30	70.27	<u>68.05</u>	66.72
AdaIGN	53.04	81.47	71.26	65.87	76.34	67.79	70.49	70.74	67.62	<u>66.79</u>
LECM	**62.90**	74.75	69.61	62.95	72.04	66.01	68.45	68.52	66.87	65.57
FrameERC	56.19	80.88	67.69	63.16	78.58	<u>69.73</u>	70.79	70.67	67.62	66.33
SEDC	<u>62.78</u>	**84.49**	<u>71.52</u>	65.71	73.63	68.09	<u>71.60</u>	<u>71.68</u>	67.43	66.16
CMDCD(Ours)	59.22	79.68	**72.86**	67.45	**79.46**	**70.14**	**72.67**	**72.69**	**68.29**	**67.08**

Table 5
Ablation study of CMDCD and WDCD module.

Method	IEMOCAP		MELD	
	Acc.	WF1	Acc.	WF1
w/o AFE	68.53	68.22	67.13	65.37
w/o CBF	69.58	69.54	67.31	66.17
w/o WDCD	71.85	71.82	67.69	66.40
w/o Wavelet	72.31	72.30	68.01	66.78
w/o Differential	72.11	72.17	68.07	66.90
w/o Conditional	72.19	72.20	68.05	66.62
CMDCD (Ours)	**72.67**	**72.69**	**68.29**	**67.08**

structure dependencies. Compared with all state-of-the-art baselines, our proposed approach CMDCD consistently achieves the best performance. For example, on the IEMOCAP dataset, CMDCD outperforms the top-performing baselines, i.e., AdaIGN and SEDC, with relative accuracy gains of 3.09% and 1.49%, respectively. Similarly, the corresponding improvements on the MELD dataset are 0.99% and 1.28%, respectively. The superior performance of our model is attributed to the multiband denoising approach grounded in wavelet differential conditional diffusion, coupled with an asymmetric fusion enhancement mechanism designed to strengthen the main modality.

It is worth noting that for specific emotion categories, CMDCD excels on Neutral, Excited, and Frustrated. While for some minority emotion categories, our method is competitive or slightly inferior to the two top-performing baselines. For example, the baseline SEDC achieves higher weighted F1 scores in the Happy and Sad categories. This is because its graph structure effectively models cross-modal global context, allowing these categories to benefit from neighboring emotional information. However, it is inferior to our model on the remaining categories, which is attributed to its vulnerability to modality-specific noise and the mutual interference among modalities.

4.6. Ablation study

To validate the effectiveness of each major component in the CMDCD framework, we conduct ablation studies by selectively removing three key modules. The details of each variant are as follows:

- **w/o AFE**: This variant discards the asymmetric fusion enhancement module, thereby disabling the model's ability to enhance modality representation by alternately focusing on each modality as the primary one while using others as auxiliaries.
- **w/o CBF**: We replace the confidence-based fusion module by conducting direct feature summation, preventing the model from adaptively integrating features from different branches based on their prediction confidence.
- **w/o WDCD**: In this variant, we remove the wavelet differential conditional diffusion module, thus eliminating the model's ability to perform multiband denoising in the frequency domain. Consequently, the model no longer selectively suppress modality-specific noise across different frequency subbands.

As summarized in Table 5, removing any of the key components of CMDCD results in noticeable performance degradation across both IEMOCAP and MELD datasets, which demonstrates that each module plays a critical role. Specifically, discarding the AFE module leads to the most pronounced performance drop. On IEMOCAP, accuracy decreases by 4.14% and weighted F1-score decreases by 4.47%. On MELD, accuracy and weighted F1-score drop by 1.16% and 1.71%, respectively. These results demonstrate the crucial role of the AFE module in mitigating the modality quality imbalance issue faced by traditional symmetric fusion strategies. By alternately emphasizing each modality as the dominant one while leveraging the others as auxiliaries, the AFE

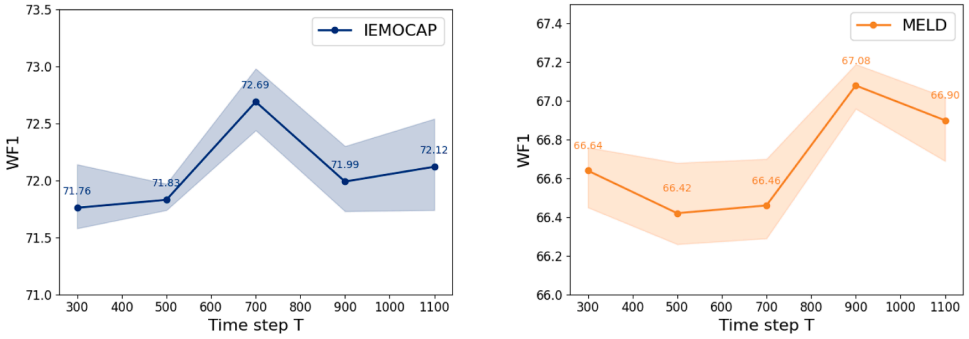


Fig. 2. Impact of diffusion steps (T) on both IEMOCAP and MELD.

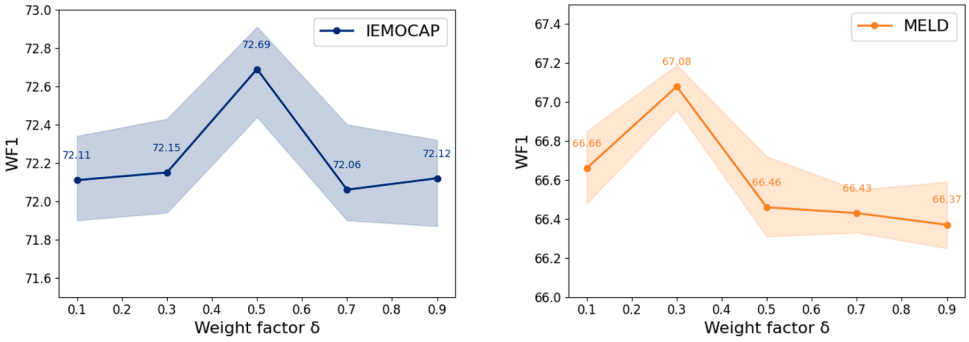


Fig. 3. Impact of weight factor (δ) on both IEMOCAP and MELD.

module facilitates the extraction of more discriminative and expressive representations, which is crucial for capturing subtle emotional cues. Similarly, removing the CBF module leads to performance degradation, which underscores the necessity of its confidence-based adaptive fusion mechanism. By adaptively balancing the contributions of asymmetric fusion-enhanced and frequency-domain denoised features via confidence-guided weighting, the CBF module effectively leverages the complementary strengths of the two components. Moreover, the removal of the WDCD module results in significant performance deterioration, which highlights its critical capability to refine semantic representations via multiband denoising, which enhances robustness by suppressing utterance-specific noise while preserving informative signals.

In addition, we also study the importance of each sub-component of the key module WDCD, including the wavelet transform, differential mechanism, and conditional guidance:

- **w/o Wavelet**: The wavelet transform is removed, and the diffusion process is performed directly in the original feature space without multiband decomposition.
- **w/o Differential**: The differential mechanism is discarded, and the model no longer separates utterance-common and utterance-specific features during the diffusion process.
- **w/o Conditional**: This variant eliminates the conditional guidance across frequency bands, rendering the diffusion process unconditional and unable to leverage inter-band complementary information.

From Table 5, we observe that ablating any of these sub-components of WDCD leads to performance degradation on both datasets, confirming that each part contributes an essential and distinct role to WDCD. Specifically, removing the wavelet transform leads to a consistent performance drop, demonstrating the importance of conducting multiband decomposition for facilitating more precise multi-scale denoising, effectively reducing utterance-specific noise across different frequency subbands. Furthermore, the ablation of the differential mechanism causes a performance degradation, confirming that distinguishing specific from common features is an essential prerequisite for precise denoising, as this mechanism is key to isolate and suppress utterance-specific noise. Finally, the removal of the conditional guidance also leads to remarkable performance degradation. This mechanism leverages cross-band signals to guide the denoising process, demonstrating its effectiveness in mitigating utterance-specific noise while preserving critical semantic features.

4.7. Sensitivity analysis

Diffusion time steps T: The hyperparameter T controls the number of iterative steps in the diffusion process, determining how extensively the model denoises the multimodal representations. We set $T \in \{300, 500, 700, 900, 1100\}$ on both datasets. The results are shown in Fig. 2. We can observe that model performance improves with an increasing number of diffusion steps T as more steps allow for finer reconstruction of the modality representations. However, when T becomes excessively large, it results in overfitting to the noise in the training data, ultimately compromising model performance. The optimal diffusion step is $T = 700$ for IEMOCAP and $T = 900$ for MELD. The higher optimal T for MELD is due to its larger scale and higher feature dimensionality, requiring a longer diffusion chain to recover detailed semantic information.

Weighting factor δ: The hyperparameter δ controls the learning degree of the weight factor in the confidence-based fusion, determining the relative contribution of each classifier. We vary $\delta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ on both IEMOCAP and MELD datasetes, the results are illustrated in Fig. 3. The performance improves with increasing δ, suggesting that a greater disparity in classifier contributions enhances the model's reliance on the most salient ones. The performance reaches its peak at $\delta = 0.5$ for IEMOCAP and $\delta = 0.3$ for MELD. Further increasing δ causes a performance decline, as excessive emphasis on the contribution differences among classifier prevents effective utilization of the complementary information from different classifiers, leading to suboptimal model performance. The distinct optimal δ values for IEMOCAP and

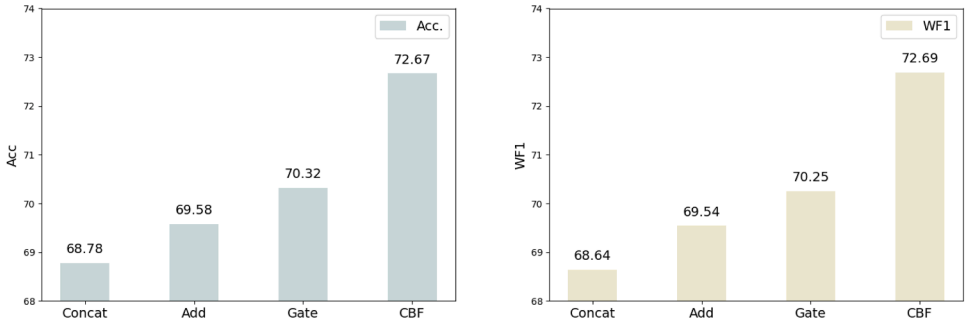


Fig. 4. Impact of four fusion method on both Acc and WF1.

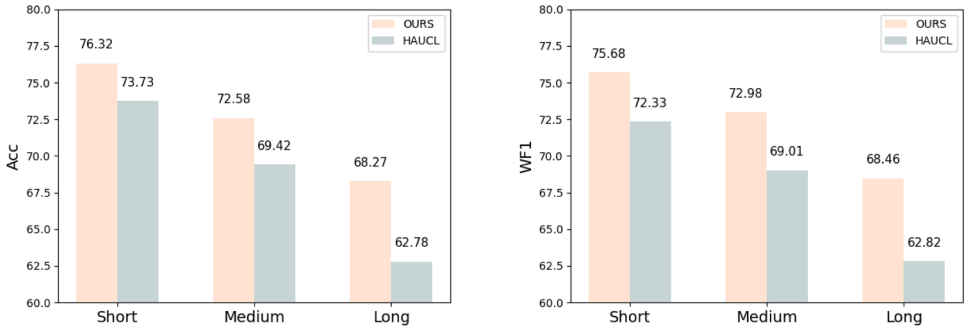
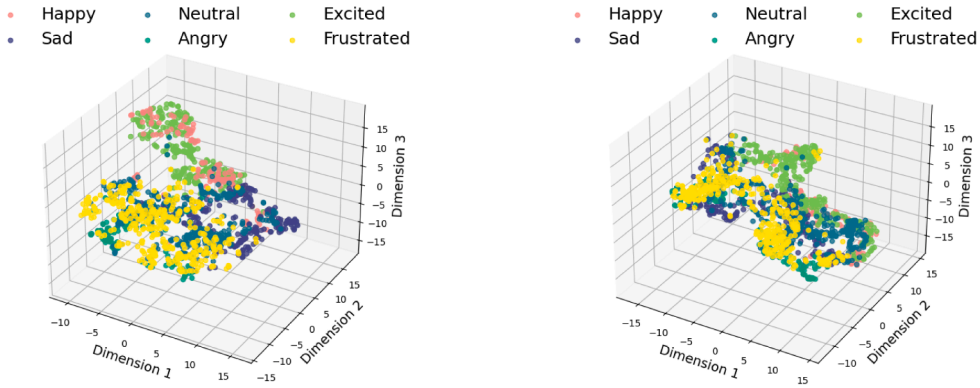


Fig. 5. Model performance under varying conversation length.



(a) HAUCL (b) CMDCD(Ours)

Fig. 6. Visualization of our proposed HAUCL and CMDCD on the IEMOCAP dataset.

MELD stem from their differing characteristics, where the former benefits from emphasizing a dominant modality and the latter's complex emotional categories necessitate a more balanced integration.

4.8. Analysis of different fusion strategies

To investigate the effectiveness of different fusion strategies, we conduct experiments on the IEMOCAP dataset using four approaches: concatenation (Concat), addition (Add), gating (Gate) [41], and our proposed confidence-based fusion (CBF).

The results are shown in Fig. 4. We can observe that CBF achieves the highest accuracy (72.67%) and weighted F1 (72.69%), consistently outperforming the other three methods. Specifically, gating performs better than concatenation and addition, indicating that introducing a learnable control mechanism helps improve fusion effect. However, its

performance gain remains limited due to the absence of explicit reliability estimation among distinct feature sources. In contrast, CBF further enhances fusion effectiveness by modeling confidence distributions, allowing model to dynamically weight the enhanced and denoised features according to their reliability. Consequently, CBF achieves a more balanced and robust integration of multimodal information, leading to a stable and superior model performance.

4.9. Sensitivity of conversation length

To investigate the sensitivity of our proposed model CMDCD under different conversation lengths, we partition the IEMOCAP dataset into three groups based on conversation lengths: *Short* for lengths less than 40, *Medium* for lengths in the range from 40 to 70, and *Long* for lengths greater than or equal to 70. The performance is compared with the HAUCL baseline [11] (AdaIGN and SEDC are omitted due to their code being publicly unavailable).

As illustrated in Fig. 5, both models exhibit a gradual decline in performance as conversation length increases. This occurs mainly owing to the fact that longer conversations usually encounter more noise and complex contextual dependencies, making the emotion recognition task more challenging. Moreover, CMDCD consistently outperforms HAUCL across all groups, and the performance gap becomes more pronounced in longer conversations. This demonstrates that our proposed model effectively learn more discriminative and robust representations, especially as complexity increases.

4.10. Visualization

To intuitively analyze the representation quality of our proposed approach CMDCD, we utilize t-SNE [42] to visualize the utterance-level embeddings obtained from CMDCD and HAUCL [11] on the IEMOCAP dataset, respectively. The results are shown in Fig. 6.

From Fig. 6, we can observe that CMDCD learns superior representations compared to HAUCL. To be specific, nodes of the same category are more cohesively grouped in CMDCD, underscoring its strength in preserving intra-class consistency. CMDCD also maintains more clear inter-cluster boundaries, reflecting its strong capability to learn discriminative representations. It is worth noting that for the neutral, excited, and frustrated emotion categories, CMDCD produces notably more compact and separable embeddings.

5. Conclusion

In this paper, we propose a Cross-Modality Multiband Differential Conditional Diffusion (CMDCD) framework for multimodal emotion recognition in conversation. The framework effectively combines asymmetric fusion enhancement and wavelet-based multiband diffusion denoising, enabling both cross-modal complementarity and fine-grained frequency-domain noise suppression. Furthermore, we employ a confidence-based adaptive fusion strategy to dynamically adjust the contribution of both enhanced and denoised representations according to their predictive confidence, thereby improving the overall interpretability and robustness of the model. Extensive experiments on the IEMOCAP and MELD benchmarks demonstrate that CMDCD consistently outperforms state-of-the-art methods in both accuracy and weighted F1-score. Ablation studies further confirm the critical role of each module and sub-component.

Despite the strong empirical performance, CMDCD still leaves room for further improvement. The current framework mainly focuses on scenarios with complete multimodal inputs, while real-world conversations may involve missing or unreliable modalities. In addition, the diffusion-based denoising process inevitably introduces extra inference cost, which may limit its deployment in ultra-low-latency applications. Moreover, although CMDCD models cross-modal interactions effectively, incorporating explicit graph-based structures could further enhance the modeling of complex and non-linear conversational dependencies. Future work will therefore explore extending CMDCD to missing-modality scenarios, developing more lightweight implementations, and integrating graph neural networks to better capture structured contextual relationships in conversations.

CRediT authorship contribution statement

Xiaofei Zhu: Writing – review & editing, Writing – original draft, Supervision, Resources, Funding acquisition, Conceptualization, Methodology; **Yang Jiang:** Writing – original draft, Software, Methodology, Investigation, Conceptualization; **Xiaoyang Liu:** Writing – review & editing, Supervision; **Yihao Zhang:** Writing – review & editing, Supervision.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (62472059), the Science and Technology Innovation Key R&D Program of Chongqing (CSTB2024TIAD-STX0027), the Chongqing Talent Plan Project, China (CSTC2024YCJH-BGZXM0022).

References

[1] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, N. Onoe, M2fNet: multimodal fusion network for emotion recognition in conversation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4652–4661.
[2] J. Li, Multimodal emotion recognition in children's online learning: emotion monitoring and intervention strategy design, Edelweiss Appl. Sci. Technol. 9 (7) (2025) 1482–1495.
[3] J. Wan, M. Woźniak, A sentiment analysis method for big social online multimodal comments based on pre-trained models, Mob. Netw. Appl. 29 (6) (2024) 1–14.
[4] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, DialogueRNN: an attentive RNN for emotion detection in conversations, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), 33, 2019, pp. 6818–6825.
[5] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, 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), 2019, pp. 154–164.
[6] M.G. Huddar, S.S. Sannakki, V.S. Rajpurohit, Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN International Journal of Interactive Multimedia and Artificial Intelligence 6 (6) (2021) 112–121.
[7] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion 91 (2023) 424–444.
[8] Z. Lian, B. Liu, J. Tao, CTNet: conversational transformer network for emotion recognition, IEEE/ACM Trans. Audio Speech Lang Process. 29 (2021) 985–1000.
[9] S. Zou, X. Huang, X. Shen, Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 5994–6003.
[10] G. Tu, T. Xie, B. Liang, H. Wang, R. Xu, Adaptive graph learning for multimodal conversational emotion detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, 38, 2024, pp. 19089–19097.
[11] Z. Yi, Z. Zhao, Z. Shen, T. Zhang, Multimodal fusion via hypergraph autoencoder and contrastive learning for emotion recognition in conversation, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 4341–4348.
[12] Z. Yang, Z. Zhang, Y. Cheng, T. Zhang, X. Wang, Semantic and emotional dual channel for emotion recognition in conversation, IEEE Trans. Affect. Comput. 16 (3) (2025) 1885–1902.
[13] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-p. Morency, S. Poria, Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis, in: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 6–15.

[14] X. Li, J. Thickstun, I. Gulrajani, P.S. Liang, T.B. Hashimoto, Diffusion-lm improves controllable text generation, Adv. Neural Inf. Process. Syst. 35 (2022) 4328–4343.
[15] D. Hu, X. Hou, L. Wei, L. Jiang, Y. Mo, MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 7037–7041.
[16] X. Zhang, Y. 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), 2023, pp. 13099–13110.
[17] J. Hu, Y. Liu, J. Zhao, Q. Jin, MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation, 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), 2021, pp. 5666–5675.
[18] F. Chen, J. Shao, S. Zhu, H.T. Shen, Multivariate, multi-frequency and multimodal: rethinking graph neural networks for emotion recognition in conversation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10761–10770.
[19] W. Lu, Z. Hu, J. Lin, L. Wang, LECM: a model leveraging emotion cause to improve real-time emotion recognition in conversations, Knowl. Based. Syst. 309 (2025) 112900.
[20] M. Li, J. Shi, L. Bai, C. Huang, Y. Jiang, K. Lu, S. Wang, E.R. Hancock, FrameERC: framelet transform based multimodal graph neural networks for emotion recognition in conversation, Pattern Recognit. 161 (2025) 111340.
[21] J. Shi, M. Li, L. Bai, F. Cao, K. Lu, J. Liang, MATCH: modality-calibrated hypergraph fusion network for conversational emotion recognition, in: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, 2025, pp. 6164–6172.
[22] K.-H. Lee, G.J. Yun, Microstructure reconstruction using diffusion-based generative models, Mech. Adv. Mater. Struct. 31 (18) (2024) 4443–4461.
[23] D. Cao, M. Chen, R. Zhang, Z. Wang, M. Huang, J. Yu, X. Jiang, Z. Fan, W. Zhang, H. Zhou, et al., SurfDock is a surface-informed diffusion generative model for reliable and accurate protein–ligand complex prediction, Nat. Methods 22 (2) (2025) 310–322.
[24] J. Zhao, W. Wang, Y. Xu, T. Sun, F. Feng, T.-S. Chua, Denoising diffusion recommender model, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 1370–1379.
[25] L. Zhang, X. Zhang, C. Li, Z. Zhou, J. Liu, F. Huang, X. Zhang, Mitigating social hazards: early detection of fake news via diffusion-guided propagation path generation, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 2842–2851.
[26] L. Yang, Y. Tian, B. Li, X. Zhang, K. Shen, Y. Tong, M. Wang, Mmada: Multimodal large diffusion language models, (2025). arXiv:2505.15809.
[27] J. Ho, A.N. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[28] Y. Song, S. Ermon, Generative modeling by estimating gradients of the data distribution, in: Adv. Neural Inf. Process. Syst. 2019, pp. 11895–11907.
[29] Y. Wang, Y. Li, Z. Cui, Incomplete multimodality-diffused emotion recognition, Adv. Neural Inf. Process. Syst. 36 (2023) 17117–17128.
[30] J. Luo, J. Wang, G. Zhou, TopicDiff: a topic-enriched diffusion approach for multimodal conversational emotion detection, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 16304–16314.
[31] X. Zhu, Y. Wang, E. Cambria, I. Rida, J.S. López, L. Cui, R. Wang, RMER-DT: robust multimodal emotion recognition in conversational contexts based on diffusion and transformers, Inf. Fusion 123 (2025) 103268.
[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: a robustly optimized bert pretraining approach, CoRR (2019).
[33] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proceedings of the 18th ACM International Conference on Multimedia, 2010, pp. 1459–1462.
[34] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.
[35] R. Gal, D.C. Hochberg, A. Bermano, D. Cohen-Or, SWAGAN: a style-based wavelet-driven generative model, ACM Trans. Graphics 40 (4) (2021) 1–11.
[36] E. Kang, W. Chang, J. Yoo, J.C. Ye, Deep convolutional framelet denosing for low-dose CT via wavelet residual network, IEEE Trans. Med. Imaging 37 (6) (2018) 1358–1369.
[37] R.S. Stanković, B.J. Falkowski, The haar wavelet transform: its status and achievements, Comput. Electr. Eng. 29 (1) (2003) 25–44.
[38] C. Zhao, W. Cai, C. Dong, C. Hu, Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8281–8291.
[39] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (2008) 335–359.
[40] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: a multimodal multi-party dataset for emotion recognition in conversations in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 527–536.
[41] P. Liu, K. Li, H. Meng, Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition, in: Proc. Interspeech 2020, 2020, pp. 379–383.
[42] M.L. van der, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605.