Multi-Condition Guided Diffusion Network for Multimodal Emotion Recognition in Conversation

Wenjin Tian[1], **Shihao Zou**[2], **Xianying Huang**[1]*
[1]Chongqing University of Technology, China
[2]Huazhong University of Science and Technology, China
wldsj_cqut@163.com

Abstract

Emotion recognition in conversation (ERC) involves identifying emotional labels associated with utterances within a conversation, a task that is essential for developing empathetic robots. Current research emphasizes contextual factors, the speaker's influence, and extracting complementary information across different modalities. However, it often overlooks the cross-modal noise at the semantic level and the redundant information brought by the features themselves. This study introduces a diffusion-based approach designed to effectively address the challenges posed by redundant information and unexpected noise while robustly capturing shared semantics, thus facilitating the learning of compact and representative features from multimodal data. Specifically, we present the Multi-Condition Guided Diffusion Network (McDiff). McDiff employs a modal prior knowledge extraction strategy to derive the prior distribution for each modality, thereby enhancing the regional attention of each modality and applying the generated prior distribution at each diffusion step. Furthermore, we propose a method to learn the mutual information of each modality through a specific objective constraints approach prior to the forward process, which aims to improve inter-modal interaction and mitigate the effects of noise and redundancy. Comprehensive experiments conducted on two multimodal datasets, IEMOCAP and MELD, demonstrate that McDiff significantly surpasses existing state-of-the-art methodologies, thereby affirming the generalizability and efficacy of the proposed model.

1 Introduction

The objective of Emotion Recognition in Conversation (ERC) is to accurately identify the emotions expressed by each participant in a conversation. This research has applications across various domains, including intelligent customer service (Han et al., 2020), human-computer interaction (Li et al., 2022), and social media analysis (Wang et al., 2023). Unlike traditional emotion recognition, which typically focuses on isolated utterances, the emotional dynamics within a conversation are influenced not only by the semantics of the utterances but also by contextual factors and the emotional interplay between speakers. Consequently, current research emphasizes the development of methodologies to extract contextual and speaker-specific information. Notable advancements in this field have been achieved through models utilizing gated recurrent units (Hazarika et al., 2018), graph neural networks (Ghosal et al., 2019; Joshi et al., 2022), and Transformer architectures (Zhang and Li, 2023; Qiu et al., 2023).

Moreover, multimodal information plays a complementary role and can significantly enhance ERC performance. Most contemporary multimodal ERC models focus on cross-modal interaction and information fusion. For example, MM-DFN (Zhang and Li, 2023) minimizes redundancy and strengthens modal complementarity by dynamically capturing contextual elements. Similarly, Joyful (Li et al., 2023) promotes deep interaction and the fusion of global context with unimodal features, while TelME (Yun et al., 2024) employs cross-modal knowledge distillation to enhance the performance of weaker modalities.

However, these approaches encounter significant challenges in extracting high-level semantics when addressing the same object across different modalities, primarily due to inconsistencies or misalignments in semantic information. For instance, while an image and its corresponding textual description may convey similar content, the image may include details that are absent from the text, or the text may emphasize aspects not represented in the image. Furthermore, these methods often overlook the redundancy inherent in multimodal data features, as different modalities may convey highly overlapping

*Corresponding author

Findings of the Association for Computational Linguistics: NAACL 2025, pages 3215–3227
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

core information. For example, both images and text may express emotions along the same dimension; however, differences in semantic space can lead to redundancy issues that hinder the effectiveness of ERC. Therefore, effectively leveraging the valuable information present in utterances to mitigate the effects of noise and redundancy is crucial for enhancing ERC performance.

Recently, the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020) has demonstrated significant success in image generation, synthesis tasks, and text generation by iteratively refining the quality of images. However, its potential for advanced semantic applications remains largely underexplored. Given its ability to gradually generate complex semantic structures, DDPM can construct high-level semantic representations from low-level features, achieving semantic alignment and a deeper understanding of multimodal data through a conditional generation mechanism. Therefore, it holds considerable promise for multimodal emotion recognition.

To address these challenges, this paper introduces a McDiff method designed to resolve the multimodal ERC issue. We implement a multi-conditional guided diffusion process to mitigate the noise resulting from high-level semantic misalignment between modalities. By incorporating modal priors and initial modal embeddings that encapsulate contextual and situational information as conditions, we enhance the collaborative functioning of different modalities, thereby reducing the adverse effects of high-level semantic inconsistencies. The McDiff method is organized into three stages: Modal Prior Knowledge Extraction, Specific Objective Constraint, and Multi-Conditional Diffusion. McDiff derives the prior distribution for each modality by extracting contextual information and employing cross-modal interaction techniques. Subsequently, the prior distribution of each modality is utilized in the specific objective constraint stage to reduce inherent redundancy in the features and enhance their generalization capabilities. Finally, this prior distribution is applied during the forward sampling process, where the prior distributions of the three modalities and the initial modal embeddings serve as conditional priors to guide the reverse process. This approach achieves distribution consistency of the fused features and facilitates semantic disambiguation. We conducted a series of experiments on two publicly available benchmark multimodal datasets, IEMOCAP and MELD, with results consistently indicating that McDiff significantly outperforms various existing multimodal ERC methods.

The primary contributions of this article are as follows:

- We propose a novel diffusion-based model for multimodal emotion recognition classification. To our knowledge, this is the first instance of a diffusion-based ERC model, which effectively mitigates unexpected noise generated during the interaction process.
- We introduce modal prior knowledge to guide the reverse process, utilizing contextual information and unimodal prior distributions generated through cross-modal interactions to adjust the diffusion steps. By conducting the diffusion process at a high semantic level across modalities, our method demonstrates fine-grained discrimination capabilities.
- We employ specific objective constraint methods to learn the mutual information within the latent space of each modality. This approach reduces feature redundancy and enables the network to model robust feature representations that are shared across multiple modalities and sessions.
- We conducted extensive experiments on two publicly available benchmark multimodal datasets, IEMOCAP and MELD. The results demonstrate that our proposed McDiff method outperforms all existing state-of-the-art baseline models in terms of effectiveness and superiority.

2 Related Work

2.1 Emotion Recognition in Conversation

The rapid growth of social media has highlighted ERC in sentiment analysis. ERC methods are divided into text-based and multimodal approaches. Text-based methods emphasize context modeling and speaker relationships, with recent research enhancing conversation-level understanding through pre-trained language models. Graph-based methods and those using common sense knowledge have also emerged (Ghosal et al., 2019). Given the complexity of emotion recognition, multimodal information is essential, providing diverse emotional cues. Significant progress in multimodal methods includes Joyful (Li et al., 2023), which introduced a fusion mechanism for deep interaction between global context and unimodal features, and a graph contrastive learning framework for better representation of emotional samples. To address encoding differences, researchers projected multimodal features into a shared subspace with a trainable basis matrix. TelME (Yun et al., 2024) applied knowledge distillation to improve cross-modal performance by transferring information from a language model to a non-language model, using a student network to support the teacher's mobile fusion method.

Recent investigations into multimodal ERC have primarily concentrated on contextual factors and cross-modal interactions. In contrast to these methodologies, our research emphasizes the challenges posed by noise and redundant information at the semantic level, drawing upon foundational research to develop the Multi-Conditional Guided Diffusion Network (McDiff). This network is designed to facilitate the guidance of semantic information within the features through a set of a priori conditions, enabling a systematic extraction of sentiment cues.

2.2 DDPM

Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) have attracted considerable interest within the domain of generative modeling in recent years. The core principle of DDPM involves the progressive addition of noise to the data until it is entirely converted into random noise, followed by the reconstruction of the original data through a methodical denoising process. This methodology can be described as a Markov chain, where each iteration involves a slight modification dictated by a Gaussian distribution. In terms of practical application, DDPM consists of two main components: the forward process and the reverse process. The forward process systematically introduces noise to the data, and its mathematical formulation is as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \qquad (1)$$

where, β_t denotes the predetermined parameter for noise scheduling, while x_t and x_{t-1} signify the states of the data at time steps t and $t-1$, respectively. Additionally, $\mathcal{N}(\cdot)$ indicates that x_t adheres to a normal distribution. The reverse process initiates from a state of pure noise and progressively reduces the noise, as represented by the following formula:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I}) \qquad (2)$$

where, $\mu_\theta(x_t, t)$ signifies the anticipated mean for inferring x_{t-1} from x_t, as determined by the parameterized model θ.The expression $\sigma_t^2\mathbf{I}$ represents the covariance matrix employed for noise modeling, with σ_t^2 regulating the magnitude of the noise, while $\mathbf{I}$ denotes the identity matrix.

The DDPM training process minimizes the discrepancy between the reverse process and the actual data distribution. Its loss function includes reconstruction loss, which keeps denoised data close to the original, and Kullback-Leibler (KL) divergence loss, which reduces the difference between the predicted and true noise distributions.

Diffusion models have been widely adapted across various fields. For instance, (Ho et al., 2020) demonstrated that DDPM excels in high-quality image generation, often surpassing Generative Adversarial Networks (GANs). The framework has also been effectively used in audio generation (Kong et al., 2021), showcasing its potential in speech synthesis. In natural language processing, researchers have applied the DDPM framework to text data (Gong et al., 2023).

Recent advancements in diffusion models include Nichol and Dhariwal's improved noise scheduling strategy (Nichol and Dhariwal, 2021), which enhances generation quality, and Salimans and Ho's accelerated sampling method (Salimans and Ho, 2022), which improves sampling efficiency.

3 Problem Formulation

A conversation consists of a sequence of utterances $U = \{u_1, u_2, \ldots, u_N\}$ and M speakers $P = \{p_1, p_2, \ldots, p_M\}$. Each utterance u_i contains n_i tokens $\{w_{i1}, w_{i2}, \ldots, w_{in_i}\}$ and is spoken by speaker $p_{\phi(u_i)}$. The N utterances correspond to N emotion labels $\{y_1, y_2, \ldots, y_N\}$. The aim of conversation emotion recognition is to analyze the conversation and classify the emotional tone of each utterance. Each utterance includes data from three modalities: Audio (a), Visual (v), and Text (t), represented as $u_i = [u_i^a, u_i^v, u_i^t]$, where $u_i^a \in \mathbb{R}^{d_a}$, $u_i^v \in \mathbb{R}^{d_v}$, and $u_i^t \in \mathbb{R}^{d_t}$. Here, $d_{(\cdot)}$ denotes the dimension of the features.

4 Proposed Model

Figure 1 provides an overview of the McDiff model we proposed. After obtaining unimodal features at the utterance level, the multi-conditional diffusion model comprises two modules: the Modal

Prior Knowledge Extraction (MPKE), which extracts prior knowledge from each modality, and the Multi-Conditional Diffusion (MCD), which learns guided by multiple conditional priors. Additionally, we introduce a specific objective constraint method and design a loss function to capture the mutual information in the latent space between modalities.

4.1 Modal Prior Knowledge Extraction

We consider three modalities: audio, visual, and text. First, we employ three independent one-dimensional temporal convolution layers to aggregate sequential information and extract low-level multimodal features: $X_m \in \mathbb{R}^{d_m}$, where $m \in \{a, v, t\}$ denotes the different modalities. After this initial encoding, each modality retains the temporal dimension of the input, allowing us to address both aligned and unaligned cases simultaneously. Furthermore, all modalities are scaled to the same feature dimension, specifically $d_t = d_v = d_a = d$. To optimize the extraction of emotional cues from a single modality, we propose a method that distinguishes between primary and secondary modalities. We alternately designate the modality requiring processing as the primary modality, while the remaining modalities provide rich semantic information to support the current primary modality. When the text modality is treated as the primary modality, it is represented as follows:

$$\begin{aligned} H_t^a &= att(X_t, X_a, X_a) \\ H_t^v &= att(X_t, X_v, X_v) \\ H_t &= att(H_t^a, H_t^v, X_t) \end{aligned} \tag{3}$$

where $att(Q, K, V)$ denotes the attention mechanism, $H_t \in \mathbb{R}^{N \times d}$. The operations for extracting emotional cues from speech and images are analogous to those used for text; consequently, we can derive $H_a \in \mathbb{R}^{N \times d}$ and $H_v \in \mathbb{R}^{N \times d}$. To capture speaker information within the utterance sequence, we utilize speaker embeddings to enhance the modal representations. In the conversation, the speaker P_j is represented as a vector. To facilitate more effective information transmission and achieve gradient control, speaker features can be integrated with modal features and input into a normalization layer. This method optimizes the model's training process while preserving interactions among different features, thereby enhancing its capacity to represent complex characteristics. The specific formula is as follows:

$$\begin{aligned} &P_j^{'} = W_{emb}O(p_j) \ \in \mathbb{R}^{3d}\ , j = 1, 2, \cdots, M, \\ &P^{'} = [P_{\phi(u_1)}^{'}, P_{\phi(u_2)}^{'}, \cdots, P_{\phi(u_N)}^{'}] \\ &[H_t^{'}; H_a{}^{'}; H_v{}^{'}] = Norm([X_t; X_a; X_v] + P^{'}) \\ &[H_t^{''}; H_a^{''}; H_v^{''}] = Norm(FFN([H_t^{'}; H_a{}^{'}; H_v{}^{'}]) \\ &\ + [H_t^{'}; H_a{}^{'}; H_v{}^{'}]) \\ &\hat{y}_m = cls(selfatt(H_m^{''}))\ , m \in (A, T, V) \end{aligned} \tag{4}$$

where, let M denote the total number of speakers, while $W_{emb} \in \mathbb{R}^{3d \times M}$ represents a trainable matrix for speaker embeddings. The notation $O(p_j) \in \mathbb{R}^M$ indicates the mapping of speaker p_j to a one-hot encoded vector, and $P' \in \mathbb{R}^{3d \times N}$ signifies the feature representation of speakers across all utterances within a conversation. The notation $[;;]$ is used to indicate a concatenation operation, $selfatt(\cdot)$ represents the self-attention mechanism, and $cls(\cdot)$ denotes the emotion classifier. Ultimately, the prior distribution $\hat{y}_m$ for each modality is obtained.

4.2 Specific Objective Constraints

To mitigate the impact of redundancy within each modality and enhance the complementarity between modalities, this paper introduces a method with specific objective constraints to learn the distributional characteristics of each modality, thereby maximizing the mutual information between modalities. Considering the stochastic nature of feature noise and the fact that real-world data adheres to a specific distribution, we adopt a weighted summation approach to reduce the effect of noise. Consequently, we take the weighted average of the prior distributions of the three modalities as the final prior distribution y_{prior}, assigning a weight of 1 to each modality. Following this, we use the following specific formula to compute the Kullback-Leibler (KL) divergence between the prior distributions of the three modalities and y_{prior}:

$$\begin{aligned} D_{KL}(p \| \, q) &= \sum_{i=1}^{N} p(i) \log \frac{p(i)}{q(i)} \\ y_{prior} &= mean(\hat{y}_a + \hat{y}_v + \hat{y}_t) \\ \mathcal{L}_{SOC} &= \sum_{m \in \{t,a,v\}} D_{KL}(y_{prior} \, || \hat{y}_m) \end{aligned} \tag{5}$$

where $p(i)$ denotes the target distribution and $q(i)$ signifies the approximate distribution. Additionally, $\hat{y}_a$, $\hat{y}_v$, and $\hat{y}_t$ represent the prior distributions

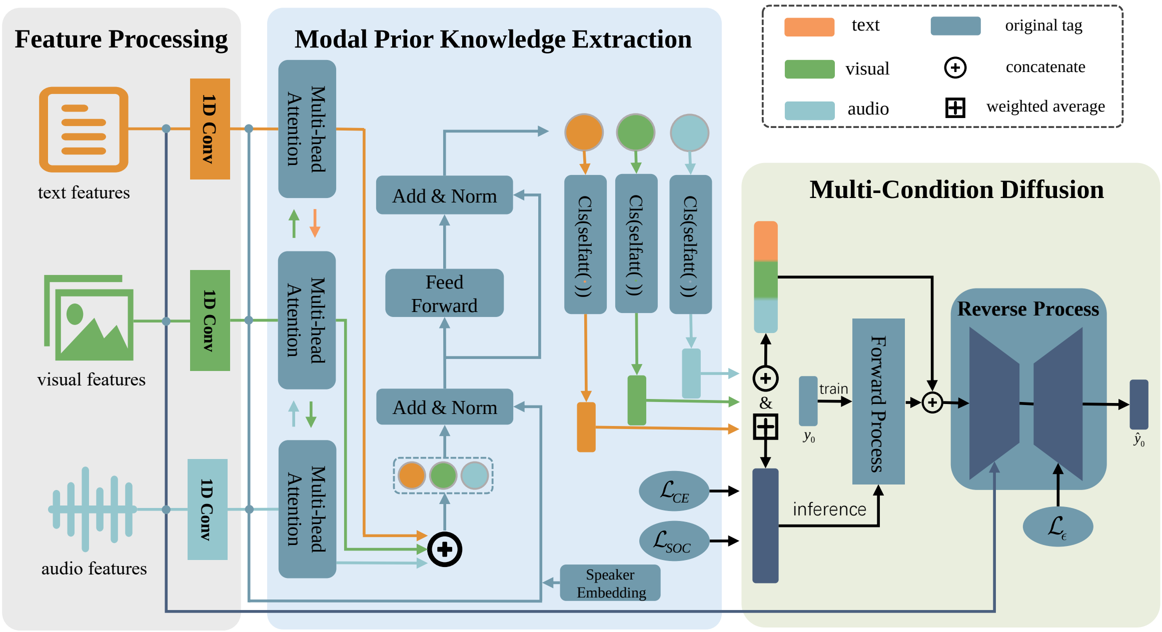

Figure 1: McDiff Model Framework Diagram.

associated with the three modalities. Typically, the cross-entropy loss and the model architecture are capable of capturing complementary information from these three priors. However, the specific objective constraints method preserves the mutual information between each prior distribution and the target distribution. This approach enhances the network's ability to effectively model shared features across various conditions, thereby facilitating more rapid and stable convergence.

4.3 Multi-Condition Diffusion

In most conditional diffusion models, the conditional prior is usually represented by a single piece of information. Discrepancies arising from multiple modalities, along with weak alignment or misalignment between these modalities, lead to unexpected noise. Additionally, the redundancy inherent in the features themselves further complicates the process of modality fusion. To address these challenges, we have developed a multi-conditional guided diffusion model that integrates the prior distributions of three modalities at the beginning of the sampling process for the noise variable y_t:

$$y_t = \sqrt{\bar{\alpha}_t} y_0 + (1 - \sqrt{\bar{\alpha}_t})(\hat{y}_a + \hat{y}_v + \hat{y}_t) \\ + \sqrt{1 - \bar{\alpha}_t} \varepsilon \tag{6}$$

where, we define $\bar{\alpha}_t = \prod_t \alpha_t$, ε follows a standard normal distribution $N(0, 1)$, and we define $\alpha_t = 1 - \beta_t$, where β_t is a parameter that controls the amount of noise added at each time step. Subsequently, we input the noise variable y_t along with the concatenated vector of the triple conditional prior into the denoising model ε_θ to estimate the noise distribution, which is expressed as:

$$\varepsilon_\theta(X_m, \hat{y}_a, \hat{y}_l, \hat{y}_v, y_t, t) \\ = D(E([\hat{y}_a; \hat{y}_l; \hat{y}_v; y_t], X_m, t), t) \tag{7}$$

where, $[\cdot]$ denotes a series operation, while $E(\cdot)$ and $D(\cdot)$ represent the encoder and decoder of the UNet, respectively. It is important to note that the initial modal feature embedding X_m is further integrated with the original embedding during the diffusion process as a condition. This integration enables the model to concentrate on high-level semantics, resulting in more robust feature representations. In the forward process, our objective is to minimize the noise estimation loss $\mathcal{L}_\epsilon$:

$$\mathcal{L}_\epsilon = ||\epsilon - \epsilon_\theta(X_m, \hat{y}_a, \hat{y}_l, \hat{y}_v, y_t, t)|| \tag{8}$$

4.4 Loss Function

In the training process of McDiff, we utilize standard cross-entropy along with specific objective constraint methods as the loss function. Furthermore, we define $\mathcal{L}_\epsilon$ as the loss function for the denoising network:

$$\begin{aligned} \mathcal{L}_{CE}^{m} &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j}\log(\hat{y}_{mi,j}) \\ \mathcal{L}_{MPKE} &= \mathcal{L}_{SOC} + \sum_{m\in\{t,a,v\}} \mathcal{L}_{CE}^{m} \end{aligned} \tag{9}$$

where N represents the total number of conversations, C denotes the number of utterances in conversation i, $\hat{y}_{mi,j}$ indicates the probability distribution of the predicted emotional label for utterance j in conversation i for modality m, and $y_{i,j}$ is the predicted category label for utterance j in conversation i. We use the Adam optimizer (Kingma and Ba, 2015) from the stochastic gradient descent algorithm to train our network model.

5 Experiments

We evaluated the effectiveness of the model using two established benchmark datasets: IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019). During the evaluation process, we first discussed the experimental results of the proposed model in comparison to the baseline model and conducted an ablation study to explore the contributions of MCD and SOC. Next, we performed a parameter analysis to investigate the experimental results and their performance trends. To further validate the model's effectiveness, we also conducted case studies and assessed the performance of the diffusion reverse process in this task through visualization experiments (for details, see Appendices A, B). For comprehensive information regarding the baseline models, feature extraction, datasets, and their implementation details, please refer to Appendices E, C, D, and F.

5.1 Comparison with Other Baseline Models

Table 2 shows the experimental results of our McDiff model against baseline models on the IEMOCAP and MELD datasets. Some data cells are missing as certain models provide only overall averages or do not use the specified evaluation metrics. Most models use graph-based structures for context propagation and cross-modal interaction, but their performance is often worse than RNN- or GRU-based baseline models. This suggests that assimilating extensive contextual information complicates feature alignment, resulting in noise and redundancy.

The CMCF-SRNet model employs a cross-modal local constraint Transformer for multimodal interaction and a graph-based semantic refinement network to enhance high-level semantics, improving performance. In contrast, TelME features a knowledge distillation module that enriches emotional cues by optimizing weaker modalities, addressing the limitations of nonverbal modes. McDiff, on the other hand, adeptly integrates modality-agnostic information and modifies distributions through a multi-condition guided reverse process, effectively mitigating high-level semantic noise and feature redundancy. As a result, McDiff outperforms all baseline models in terms of accuracy and weighted F1 score in overall multimodal emotion recognition.

As shown in Table 1, the McDiff model outperforms baselines in most emotion classification tasks on the IEMOCAP dataset but struggles with "Sad" and "Frustrated" categories due to overlapping acoustic features, particularly in speech prosody. Despite this, it maintains high accuracy.

In the MELD dataset, Figure 2 reveals class imbalance, with neutral samples dominating and reducing the recognition rate of minority classes such as "fear" and "disgust". It is worth noting that the "happy" and "sad" labels show different performances across models (Table 1), which may be due to the limited labels and complexity of emotions, resulting in the lack of generalizable feature learning for these labels.

5.2 Ablation Study

To investigate the contributions of the two modules proposed in McDiff, we systematically eliminated the primary components of McDiff and assessed their effects on model performance. McDiff is pri-

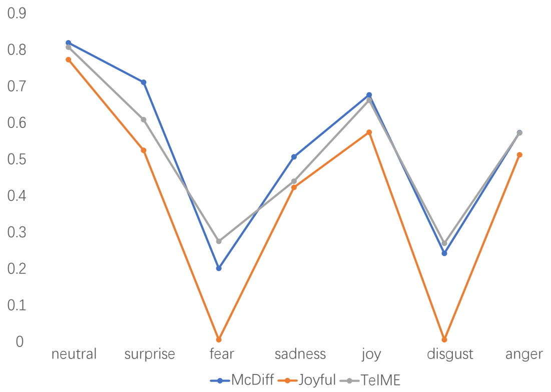

Figure 2: Comparison of McDiff with two baseline models, Joyful and TelME, on the MELD dataset across each emotion category.

Table 1: Comparison of $WF1$ for each emotion category between McDiff and other baseline models on the IEMOCAP test set.

	IEMOCAP					
	Happy	Sad	Neutral	Angry	Excited	Frustrated
ICON	32.80	74.40	60.60	68.20	68.40	66.20
DialogueGCN	42.75	**84.54**	63.54	64.19	63.08	66.99
MMGCN	42.34	78.67	61.37	69.00	74.33	62.32
MM-DFN	42.22	78.98	66.42	69.77	75.56	66.33
COGMEN	51.90	81.70	68.60	66.00	75.30	58.20
Joyful	60.94	84.42	68.24	69.95	73.54	67.55
CMCF-SRNet	52.20	80.90	68.80	70.30	76.70	61.60
CORECT	59.30	80.53	66.94	69.59	72.69	**68.50**
TelME	-	-	-	-	-	-
McDiff	**62.69**	83.47	**72.75**	**70.33**	**78.36**	68.19

Table 2: Experimental results on the IEMOCAP and MELD datasets. The blanks indicate instances where the baseline is not open-sourced or where a specific evaluation metric was not employed. The best results are highlighted in bold.

	IEMOCAP		MELD	
	ACC	WF1	ACC	WF1
ICON	64.00	63.50	-	-
DialogueGCN	63.22	62.89	60.31	56.36
MMGCN	66.06	65.65	61.26	57.97
MM-DFN	68.21	68.18	62.49	59.46
COGMEN	68.20	67.60	-	-
Joyful	70.55	71.03	62.53	61.77
CMCF-SRNet	70.50	69.60	62.30	-
CORECT	69.93	70.02	-	-
TelME	70.48	-	67.37	-
McDiff	**72.77**	**73.19**	**67.94**	**68.78**

marily composed of three modules: Modal Prior Knowledge Extraction (MPKE), Specific Objective Constraint (SOC), and Multi-Condition Guided Diffusion (MCD). In this analysis, we focus exclusively on the latter two modules, considering MPKE as a fundamental classifier that can operate as a complete classifier by executing a weighted average of the three generated modal priors. The following settings are considered in our study:

- basic: We have eliminated all diffusion operations and specific objective constraint losses from the network, which means that the basic model utilizes the most fundamental classifier.
- C1: We applied the MCD to the basic model, thereby constructing an additional baseline network.
- C2: Further apply the method of SOC to the diffusion process to construct the baseline network.

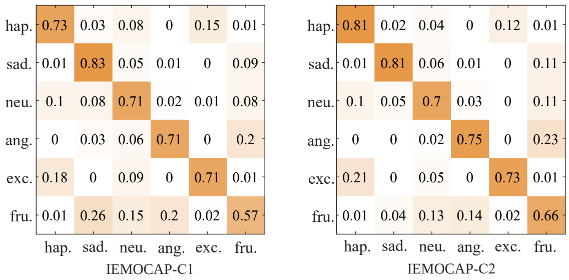

Figure 3: Confusion matrix of each ablation group in the IEMOCAP dataset.

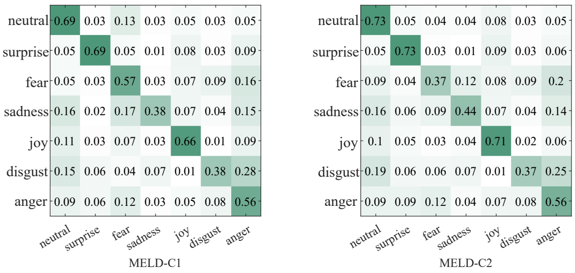

Figure 4: Confusion matrix of each ablation group in the MELD dataset.

We conducted a series of experiments to evaluate the efficacy of the proposed network components and developed three baseline networks based on our research methodology. The first network was constructed by removing all diffusion operations and SOC losses. Subsequently, we introduced the

Table 3: Efficacy of each module in our McDiff system.

			IEMOCAP		MELD	
	MCD	SOC	ACC	WF1	ACCC	WF1
Basic	-	-	0.679	0.681	0.652	0.663
C1	✓	-	0.698	0.690	0.667	0.679
C2	✓	✓	0.728	0.732	0.679	0.688

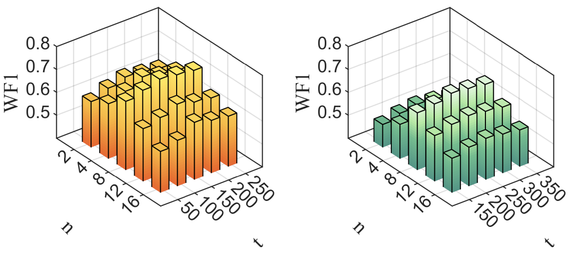

Figure 5: Presents a three-dimensional bar chart illustrating the results of parameter sensitivity experiments conducted on the IEMOCAP dataset (left) and the MELD dataset (right).

multi-condition diffusion process, resulting in the "C2" network, which incorporates constraint loss. As illustrated in Figures 3 and 4, "C2" significantly outperforms "C1" in terms of the F1 score across individual categories. Table 3 presents the experimental results for accuracy and weighted F1 scores on the IEMOCAP and MELD datasets. Compared to the baseline model, "C1" improved accuracy by 0.019 and weighted F1 by 0.009, indicating the effectiveness of the multi-condition guided diffusion mechanism in extracting discriminative features for multimodal emotion classification. Furthermore, "C2" surpassed the baseline model in both accuracy and weighted F1, demonstrating that the combination of our proposed guiding strategy and constraint method significantly enhances model performance. Additionally, "C2" also exceeded "C1" in accuracy and weighted F1, further suggesting that the introduction of constraint loss prior to the diffusion process effectively improves the model's performance.

5.3 Parameter Analysis

We conducted experiments to examine the effects of varying the number of attention heads (n) and the diffusion time step (t) on model performance. Figure 5 shows the results across different configurations. For the IEMOCAP dataset, n varied from 4 to 16 and t from 50 to 250, while for the MELD dataset, n remained the same, and t ranged from 150 to 350. We selected a diffusion step size lower than typically used in literature, as our findings indicated that exceeding 2000 steps significantly reduced the model's restoration capabilities. Conversely, lower step sizes improved performance, likely due to the sensitivity of high-level semantic cues and dataset limitations. Optimal performance for McDiff on the IEMOCAP dataset was achieved with $n = 8$ and $t = 150$, while satisfactory performance on the MELD dataset was attained with $n = 8$ and $t = 250$.

6 Conclusion

This article introduces a novel multimodal emotion recognition approach known as McDiff. The fundamental premise of this model is the incorporation of a multi-condition guidance strategy into the conventional diffusion model (DDPM), alongside the application of specific objective constraint losses prior to the diffusion process to improve classification efficacy. Through a series of ablation experiments, we elucidate the significance of each component within McDiff. In comparison to existing literature on ERC, our experimental findings on multimodal classification datasets reveal that McDiff surpasses current state-of-the-art methodologies in terms of performance.

Limitation

McDiff faces significant challenges related to high computational complexity and the complexities of cross-modal collaborative processing. The generative process inherent in diffusion models requires multiple sampling iterations, leading to considerable consumption of computational resources. Additionally, McDiff must effectively integrate information from various modalities, including audio, visual, and textual data, throughout the diffusion process. Experimental results suggest that the integration of information across different modalities presents difficulties due to potential asynchrony or inconsistency. Although the model employs additive and concatenation techniques, further optimization and the incorporation of additional technologies are necessary to improve overall performance. Future research will focus on accelerating algorithm development and enhancing cross-modal collaborative processing.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62472059). We would like to express our gratitude to our colleagues for their valuable insights and to the anonymous reviewers for their constructive feedback.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4):335–359.

Florian Eyben, Martin Wöllmer, and Björn W. Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, pages 1459–1462. ACM.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. 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)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. Diffuseq: Sequence to sequence text generation with diffusion models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Wenjing Han, Tao Jiang, Yan Li, Björn W. Schuller, and Huabin Ruan. 2020. Ordinal learning for emotion recognition in customer service calls. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 6494–6498. IEEE.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. 2022. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7037–7041. IEEE.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. 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)*, pages 5666–5675, Online. Association for Computational Linguistics.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, Seattle, United States. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. 2023. Joyful: Joint modality fusion and graph contrastive learning for multimoda emotion recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16051–16069. Association for Computational Linguistics.

Sha Li, Mahdi Namazifar, Di Jin, Mohit Bansal, Heng Ji, Yang Liu, and Dilek Hakkani-Tur. 2022. Enhancing knowledge selection for grounded dialogues via document semantic graphs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2810–2823, Seattle, United States. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Leland McInnes and John Healy. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426.

Cam-Van Thi Nguyen, Anh-Tuan Mai, The-Son Le, Hai-Dang Kieu, and Duc-Trong Le. 2023. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15154–15167. Association for Computational Linguistics.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Shuwen Qiu, Nitesh Sekhar, and Prateek Singhal. 2023. Topic and style-aware transformer for multimodal emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2074–2082. Association for Computational Linguistics.

Tim Salimans and Jonathan Ho. 2022. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yusong Wang, Dongyuan Li, Kotaro Funakoshi, and Manabu Okumura. 2023. EMP: emotion-guided multi-modal fusion and contrastive learning for personality traits recognition. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, ICMR 2023, Thessaloniki, Greece, June 12-15, 2023*, pages 243–252. ACM.

Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. 2024. Telme: Teacher-leading multimodal fusion network for emotion recognition in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 82–95. Association for Computational Linguistics.

Xiaoheng Zhang and Yang Li. 2023. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13099–13110. Association for Computational Linguistics.

A Visualization of the Diffusion Process

To illustrate the inverse process of diffusion guided by multiple conditions, we use the UMAP (McInnes and Healy, 2018) tool to visualize the denoised feature embeddings at consecutive time steps. Figure 6 shows the results of this process on two datasets. As the time steps advance, the denoising diffusion model gradually removes noise from the feature representation, making the class distribution in the Gaussian distribution clearer, further verifying the effectiveness of the McDiff model structure. The total number of time steps required for inference depends on the complexity of the dataset.

B Case Study

Figure 7 shows a conversation excerpt from the MELD dataset. Unlike the IEMOCAP dataset, MELD has fewer utterances per conversation, complicating the emotion recognition task. This conversation features nine utterances between two speakers, highlighting emotional transitions. Speaker S1 feels neglected, leading to anger and demands on Speaker S2, which in turn provokes S2's anger. This results in a trust crisis, with S1 questioning S2's loyalty. Ultimately, S2 decides to end the relationship, leading to sadness. Our findings reveal that when a speaker's emotions transition, our

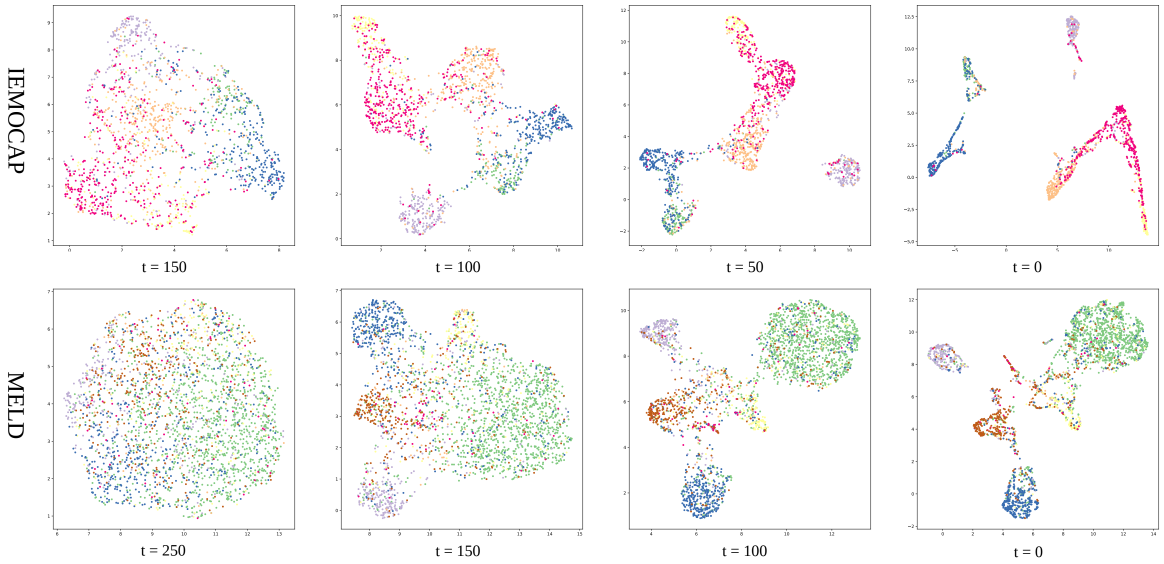

Figure 6: Shows the denoised feature embeddings from the diffusion reverse process during inference across two datasets. The variable t indicates the current time step, and as it progresses, noise is reduced, leading to a clearer class distribution. For better visibility, refer to the last column, which can be enlarged.

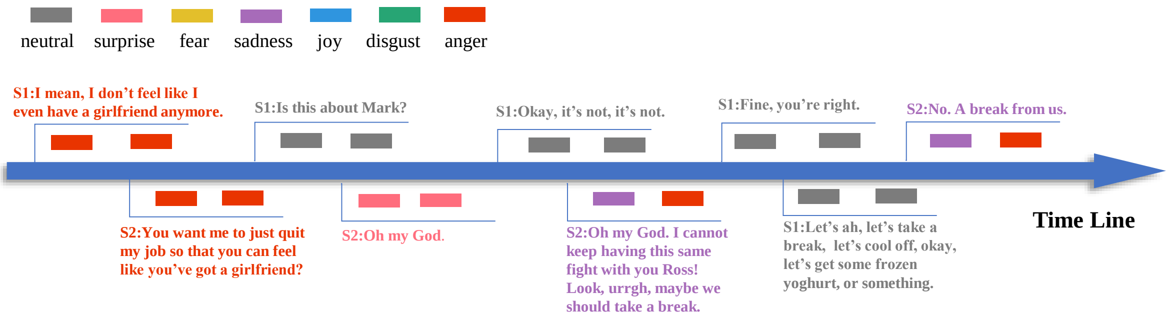
neutral
surprise
fear
sadness
joy
disgust
anger
S1:I mean, I don't feel like I even have a girlfriend anymore.
S1:Is this about Mark?
S1:Okay, it's not, it's not.
S1:Fine, you're right.
S2:No. A break from us.
S2:You want me to just quit my job so that you can feel like you've got a girlfriend?
S2:Oh my God.
S2:Oh my God. I cannot keep having this same fight with you Ross! Look, urrgh, maybe we should take a break.
S1:Let's ah, let's take a break, let's cool off, okay, let's get some frozen yoghurt, or something.
Time Line

Figure 7: Illustrates a case analysis, where each color denotes a specific emotion. The utterance color indicates the conveyed emotion, and below it are the emotion predictions from the McDiff model (left) and the TelME model (right).

model performs better than in the Joyful category, indicating that multi-condition reinforcement enhances its understanding of emotional cues and improves performance in complex emotional dynamics.

C Feature Extraction

C.1 Text Feature

Using the RoBERTa Large model (Liu et al., 2020) to extract text features, we leverage a pre-trained architecture based on a multi-layer transformer encoder, which represents an enhancement over the original BERT model. This model has been trained on larger and more diverse datasets, enabling it to effectively learn nuanced text representations. We fine-tuned RoBERTa Large to identify emotions in conversation text, utilizing the embedding of the [CLS] token from the final layer as the text feature. The dimensionality of the extracted text features is 1024.

C.2 Audio Feature

Using openSMILE (Eyben et al., 2010) for audio feature extraction, openSMILE is a versatile toolkit specifically designed for signal processing. It offers a scriptable console application that allows users to configure modular feature extraction components. After utilizing the openSMILE toolkit, the Fully Connected (FC) layer reduces the audio feature representation of IEMOCAP to 1,582 dimensions, while MELD is reduced to 300 dimensions.

C.3 Visual Feature

Using DenseNet (Huang et al., 2017) for pre-training on the Facial Expression Recognition Plus dataset allows for the extraction of visual features. DenseNet is an efficient convolutional neural network (CNN) architecture composed of multiple dense blocks, each containing several layers. The output dimension of DenseNet is set to 342, indicating that the dimensionality of the visual feature representation is 342.

D Dataset and Evaluation Metrics

We conducted an assessment of our model's efficacy utilizing two established benchmark datasets, namely IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019). Both datasets are characterized as multimodal ERC datasets, encompassing text, speech, and visual elements. The distribution of data across the two datasets is presented in Table 4. IEMOCAP is a multimodal ERC dataset comprising conversations performed by pairs of actors based on scripted scenarios. It encompasses a total of 7,433 utterances across 151 conversations, with each utterance categorized into one of six emotional classifications: happy, sad, neutral, angry, excited, and frustrated. Conversely, MELD derives its data from the television series "Friends," containing 13,708 utterances and 1,433 conversations. Unlike the binary structure of IEMOCAP, MELD incorporates conversations featuring three or more speakers, with each utterance assigned to one of seven emotional categories: neutral, surprise, fear, sadness, joy, disgust, and anger. Following the work of previous researchers in this area, we aim to comprehensively assess the classification performance of the model for each category while considering the imbalanced classes. In the upcoming experiments, we will utilize the weighted F1 score ($WF1$) and accuracy (ACC) of various sentiment categories to evaluate the performance of McDiff. The calculation methods for $WF1$ and ACC are outlined below:

Table 4: Distribution of IEMOCAP and MELD.

Dataset	#Conversation		#Utterance	
	train+val	test	train+val	Test
IEMOCAP	120	31	5810	1623
MELD	1153	280	11098	2610

$$WF1 = \frac{\sum_{j=1}^{R} N_j * F1_j}{\sum_{j=1}^{R} N_j} \tag{10}$$

$$ACC = \frac{\sum_{j=1}^{R} N_j * Accuracy_j}{\sum_{j=1}^{R} N_j} \tag{11}$$

where, R represents the number of emotion categories in the dataset, N_j is the number of statements in the j-th emotion category, $F1_j$ represents the score for the j-th emotion category, while $Accuracy_j$ indicates the accuracy score for the j-th emotion category.

E Baseline Model

1)**ICON** (Hazarika et al., 2018) employs two Gated Recurrent Units (GRUs) to model speaker information and track emotional state transitions with an additional global GRU. It features a multi-layer memory network for overall emotional state representation but struggles in multi-speaker environments. 2) **DialogueGCN** (Ghosal et al., 2019) uses Graph Convolutional Networks (GCNs) for ERC, generating comprehensive features. Both Relational GCN (RGCN) and GCN are non-spectral models for encoding graph data. 3) **MMGCN** (Hu et al., 2021) employs a GCN framework to extract contextual information, overcoming DialogueGCN's limitations in multimodal dependencies while incorporating speaker information for emotion recognition. 4) **MM-DFN** (Hu et al., 2022) improves multimodal contextual integration through a graph-based dynamic fusion module, enhancing emotion identification in conversations. 5) **COGMEN** (Joshi et al., 2022) presents a multimodal emotion recognition system using a Contextualized Graph Neural Network that integrates local and global contexts to model complex conversational dependencies. 6) **Joyful** (Li et al., 2023) tackles the challenge of representing global and local features in multimodal emotion recognition through modality fusion, optimizing graph contrastive learning and emotion recognition together. 7) **CMCF-SRNet** (Zhang and Li, 2023) enhances multimodal interaction with a framework that integrates cross-modal local constraint context fusion and a semantic refinement module for improved ERC. 8) **CORECT** (Nguyen et al., 2023) introduces a relational temporal graph neural network that captures session-level cross-modal interactions and utterance-level temporal dependencies, emphasizing key factors represented by modalities. 9) **TelME** (Yun et al., 2024) uses cross-modal knowledge distillation to transfer information from a language model to a non-language model, enhancing the weaker modality's effectiveness and strengthening the teacher's mobile fusion approach for multimodal feature integration.

F Implementation Details

The proposed model is implemented using the PyTorch framework, with specific hyperparameter configurations. The initial learning rate is set to 1×10^{-4} for the IEMOCAP dataset and 1×10^{-5} for the MELD dataset. A batch size of 16 is utilized for both datasets. In the context of the one-dimensional convolutional layers, the input channels for the text, audio, and visual modalities in IEMOCAP are configured to 1024, 1582, and 342, respectively, corresponding to their respective feature dimensions. For the MELD dataset, these values are adjusted to 1024, 300, and 342. A dropout rate of 0.5 is applied uniformly across both datasets. The MPKE module incorporates eight attention heads. The diffusion time step t is drawn from a uniform distribution within the range $[1, T]$, and the noise is scheduled linearly with $\beta_1 = 1 \times 10^{-4}$ and $\beta_t = 0.02$. Empirical observations indicate that the total diffusion time step T is set to 150 for IEMOCAP and 250 for MELD, which is notably lower than that reported in most existing studies. Following a pre-training phase of the classification model for 10 epochs, the denoising diffusion model and the classification model are jointly trained to achieve the end-to-end McDiff for multimodal emotion recognition. All training and testing procedures are conducted on a single RTX 3090 GPU, with the total training time amounting to 8.3 hours for the IEMOCAP dataset and 10.6 hours for the MELD dataset, respectively.