



DTDiff: adaptive decoupled transformer with language-conditioned denoising learning for multimodal emotion recognition in conversation

Tingting Zhang¹ · Xiaofei Zhu¹

Received: 5 July 2025 / Revised: 19 September 2025 / Accepted: 22 September 2025

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

Abstract

Multimodal Emotion Recognition in Conversation (MERC) has attracted significant attention in recent years, and existing methods mainly rely on contextual cues and multimodal interactions to predict emotions. However, these methods often suffer from the detrimental effects of noise, including contextual interaction noise and modality-specific noise contamination, leading to suboptimal model performance. Therefore, we propose DTDiff, a noise-aware framework that tackles two types of noise: inter-utterance interaction noise through an Adaptive Residual Decoupled Transformer (ARDT), and modality-specific noise via Language-Conditioned Denoising Learning (LDL). Specifically, ARDT improves robustness by effectively filtering irrelevant contextual dependencies and enhances the representation of each modality through decoupled residual attention fusion. Meanwhile, LDL employs language-conditioned diffusion models to denoise visual and acoustic modalities and measures noise levels via gating mechanisms. Finally, we introduce Dual-Signal Alignment to further promote multimodal fusion. Experiments on IEMOCAP and MELD datasets demonstrate that DTDiff outperforms state-of-the-art methods.

Keywords Emotion recognition in conversation · Transformer · Diffusion model · Multimodal fusion

1 Introduction

Multimodal Emotion Recognition in Conversation (MERC) holds significant research value across a wide range of fields, including its applications in recommendation systems (Zheng et al., 2022), social media analysis (Wu et al., 2024), and human communication contexts

✉ Xiaofei Zhu
zxf@cqut.edu.cn

Tingting Zhang
ztt@stu.cqut.edu.cn

¹ College of Computer Science and Engineering, Chongqing University of Technology, Hongguang Avenue, Chongqing 400054, China

(Jannu & Vanambathina, 2025). Compared to methods relying exclusively on the language modality, the integration of multimodal signals including vocal tones and facial expressions allows for more comprehensive analysis of human affective states (Yun et al., 2024; Chen et al., 2023; Li et al., 2023a). This critical advantage has motivated substantial progress in MERC, particularly in contextual modeling (Joshi et al., 2022; Li et al., 2023a; Shou et al., 2024) and multimodal fusion techniques (Sun et al., 2023; Li et al., 2023b; Hu et al., 2022). However, existing methods still face many challenges.

To model contextual dependencies in conversations, graph neural networks (GNN) have become a predominant paradigm due to their capability of handling complex conversational structures (Ghosal et al., 2019; Chen et al., 2023; Shen et al., 2021; Ishiwatari et al., 2020; Nguyen et al., 2024; Shou et al., 2024; Hu et al., 2021; Tu et al., 2024; Yi et al., 2024; Gan et al., 2025; Wu et al., 2025b). For instance, DialogueGCN (Ghosal et al., 2019) utilizes graph convolutional networks (GCN) to capture intra-speaker and cross-speaker dependencies, thereby enhancing contextual information propagation. DAG-ERC (Shen et al., 2021) and MultiDAG (Nguyen et al., 2024) represent conversations as directed acyclic graphs (DAG), simulating information flow through speaker identity and positional constraints. RGAT (Ishiwatari et al., 2020) incorporates relational graph attention networks with positional encoding to strengthen sequential dependency modeling. While existing methods establish contextual interactions through speaker position encoding, they overlook whether the interacting utterances are semantically relevant. As illustrated in Fig. 1, the second and fourth utterances from the same speaker, but have minimal semantic relevance and opposing emotional polarities. However, existing methods propagate misleading information from the second to the fourth utterance, thereby introducing contextual interaction noise. Beyond graph-based methods, sequence modeling techniques such as BC-LSTM (Poria et al., 2017) and DialogueRNN (Majumder et al., 2019) extract contextual patterns using LSTM and RNN, respectively. However, such sequential models (Jiao et al., 2019) are constrained to



Fig. 1 An example of MERC from the MELD dataset. Words highlighted in red indicate emotional cues in the language modality

local temporal interactions. Many methods leverage Transformer architectures (Li et al., 2020; Zong et al., 2023; Zhang & Li, 2023; Ma et al., 2024; Jing & Zhao, 2024; Zou et al., 2023; Yun et al., 2024) to capture global information, where self-attention matrices reflect the correlation between utterances. Nevertheless, conventional Transformers suffer from two critical limitations: (1) the attention matrices are affected by contextual interaction noise, resulting in a lack of robustness of the model, and (2) there are quality discrepancies among multi-head attention matrices, which lead the low-quality attention matrices to degrade performance. For multimodal fusion, while existing methods focus on how to effectively integrate the modality information from different semantic spaces (Li et al., 2023b; Yun et al., 2024; Zou et al., 2023; Zhang et al., 2024a; Li et al., 2023c), they inadequately account for inherent modality-specific noise. According to Fig. 1, it can be observed that the fourth utterance has background noise contradicting the emotional intent of the speaker in acoustic modality, leading the model to misjudge this utterance as a positive emotion. Furthermore, language modalities often provide clearer emotional cues compared to acoustic and visual signals, as evidenced by the second, fourth and fifth utterances in Fig. 1. Therefore, compared with the language modality, there is more noise in the acoustic and visual modalities.

To address the aforementioned challenges, we propose DTDiff, a novel framework designed for MERC. DTDiff tackles two key issues: contextual interaction noise and modality-specific noise, through its two core components: the Adaptive Residual Decoupled Transformer (ARDT) and Language-Conditioned Denoising Learning (LDL). On one hand, the ARDT module (detailed in Section 3.3) introduces implicit learnable masking and explicit speaker-aware signals to dynamically adjust attention matrices. These signals effectively filter contextual interaction noise, enabling the attention scores to more accurately reflect inter-utterance semantic relevance. Furthermore, to address quality discrepancies among multi-head attention matrices, ARDT computes head-wise importance scores through cross-head interaction, thereby suppressing low-quality attention heads during fusion (detailed in Section 3.3.1). On the other hand, the LDL module (detailed in Section 3.4) employs language-conditioned diffusion processes to progressively denoise acoustic and visual modalities. The denoised features are then aligned via the Dual-Signal Alignment module (detailed in Section 3.4.2). Finally, recognizing the noise levels in different utterances and modalities, we design learnable gating weights to adaptively fuse features before and after denoising, ensuring optimal fusion under varying noise conditions (detailed in Section 3.4.3). In summary, the main contributions of this paper are summarized as follows:

1. We propose ARDT which effectively mitigates contextual interaction noise through masking signals and evaluates attention quality via interactions between attention matrices to enhance model robustness.
2. We design LDL, a language-conditioned denoising module that denoises acoustic and visual modalities, followed by feature alignment and adaptively fuse multimodal features.
3. Extensive experiments demonstrate that our framework achieves state-of-the-art performance on both IEMOCAP and MELD benchmarks, outperforming existing methods by significant margins.

2 Related work

2.1 Multimodal emotion recognition in conversation

Current approaches to Multimodal Emotion Recognition in Conversation (MERC) primarily fall into two categories: graph-based methods and Transformer-based methods. Given the sequential nature and speaker positions of utterances in a conversation, some studies (Hu et al., 2021, 2022; Tu et al., 2024; Nguyen et al., 2024; Yi et al., 2024; Wu et al., 2025b; Job et al., 2025) employ graph structures to facilitate contextual or inter-modal interactions. For instance, MultiDAG (Nguyen et al., 2024) uses directed acyclic graph network to address emotional shifts. AdaIGN (Tu et al., 2024) presents an adaptive interactive graph network with self-supervised learning to model intra- and inter-speaker dependencies. HAUCL (Yi et al., 2024) introduces a framework that dynamically adjusts hypergraph connections using a variational hypergraph autoencoder. However, these models often struggle to capture global information, which limits their performance in long conversations.

Given the Transformer's ability to directly relate elements at any position in a sequence, effectively aggregating context information by self-attention mechanism (Parisae & Nagakishore Bhavanam, 2024; Jannu & Vanambathina, 2023a), its architecture has achieved remarkable success across a wide range of fields in recent years (Jannu & Vanambathina, 2023b; Wu et al., 2025a). In the MERC task, leveraging self-attention mechanisms, Transformers effectively process and fuse cross-modal information (Xie et al., 2022; Lian et al., 2021; Zhang & Li, 2023; Ma et al., 2024; Zong et al., 2023; Jing & Zhao, 2024; Shi et al., 2025). For example, CTNet (Lian et al., 2021) develops a multimodal framework explicitly modeling intra-modal relationships and cross-modal interactions. CMCF-SRNet (Zhang & Li, 2023) integrates locality-constrained transformers with graph-enhanced semantic refinement for utterance relationship modeling. SDT (Ma et al., 2024) employs self-distilled intra- and inter-modal transformers with hierarchical gated fusion for dynamic interaction capture. However, these methods often overlook critical challenges such as contextual interaction noise interference and quality discrepancies among multi-head attention matrices.

2.2 Diffusion model

Diffusion models (Ho et al., 2020), as advanced generative frameworks, have been widely applied in image (Tang et al., 2024; Zhang et al., 2024b) and text (Li et al., 2022; Lin et al., 2025) generation. Beyond these domains, they demonstrate significant value across diverse research areas, including computer vision tasks such as image segmentation (Li et al., 2024a; Suryanto et al., 2025), object detection (Sun et al., 2025; Zhang et al., 2025), and image restoration (Chen et al., 2024; Yue et al., 2025). In natural language processing, diffusion models are increasingly adopted for data augmentation (Le et al., 2025; Luo et al., 2024) and denoising (Jiang et al., 2024; Li et al., 2024b), achieving notable performance improvements. Given their exceptional denoising capabilities and the limited exploration of conditional diffusion in MERC, we adopt conditional diffusion to denoise the acoustic and visual modalities, thereby enhancing model performance.

3 Methods

This section details our proposed framework, DTDiff, whose architecture is illustrated in Fig. 2. The framework comprises four components: (1) Modality-specific Contextual Extraction designed to capture contextual information for each modality (Section 3.2); (2) Adaptive Residual Decoupled Transformer (ARDT) that enhances modality representations through adaptive attention masking signals and head importance weights (Section 3.3); (3) Language-Conditioned Denoising Learning (LDL) that mitigates noise from acoustic and visual modality by language-conditioned diffusion (Section 3.4); and (4) the Classifier that generates emotion predictions with joint optimization (Section 3.5).

3.1 Task definition

A conversation $D = \{u_1, u_2, \dots, u_N\}$ consists of N utterances, each utilizing information from three modalities, i.e., *acoustic* (a), *visual* (v) and *language* (l). Specifically, the i -th utterance u_i is represented as $u_i = \{\mathbf{u}_i^a, \mathbf{u}_i^v, \mathbf{u}_i^l\}$, where acoustic features $\mathbf{u}_i^a \in \mathbb{R}^{d_a}$ are extracted using OpenSmile, visual features $\mathbf{u}_i^v \in \mathbb{R}^{d_v}$ through DenseNet, and language features $\mathbf{u}_i^l \in \mathbb{R}^{d_l}$ via the RoBERTa-Large model. The MERC task aims to predict the emotional state of each utterance by effectively integrating these multimodal features.

3.2 Modality-specific contextual extraction

Considering the critical role of contextual information in conversation and the proven ability of GRUs to capture long-range dependencies across input sequences (Jannu & Vanambathina, 2023c), we first employ BiGRUs to extract contextual features between adjacent utterances:

$$\tilde{\mathbf{x}}_i^m, \mathbf{h}_i^m = \text{BiGRU}^m(\mathbf{W}_0^m \mathbf{u}_i^m + \mathbf{b}_0^m, \mathbf{h}_{i-1}^m), m \in \{a, v, l\}, \tag{1}$$

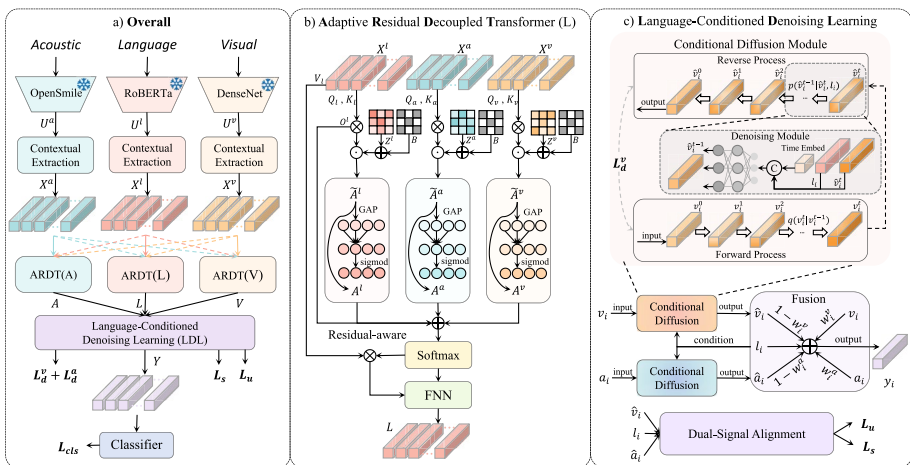


Fig. 2 An overview of DTDiff: a) The entire process of our proposed model, which includes Contextual Extraction, Adaptive Residual Decoupled Transformer (ARDT), Language-Conditioned Denoising Learning (LDL), and Classifier; b) ARDT (Language); c) LDL

where $W_0^m \in \mathbb{R}^{d_0 \times d_m}$ and $\mathbf{b}_0^m \in \mathbb{R}^{d_0}$ are learnable parameters, $\mathbf{h}_i^m \in \mathbb{R}^{d_0}$ and $\tilde{\mathbf{x}}_i^m \in \mathbb{R}^{d_0}$ denotes the hidden state and output of the BiGRU for the i -th utterance, respectively.

Furthermore, we apply DAG-ERC(Shen et al., 2021) for each modality to capture long-distance contextual dependencies between different speakers:

$$X^m = \text{DAG}^m(\tilde{X}^m), \quad (2)$$

where $X^m = \{\mathbf{x}_i^m\}_{i=1}^N \in \mathbb{R}^{N \times d}$ represents the conversation sequence in modality m .

3.3 Adaptive residual decoupled transformer

Since the multi-head attention mechanism in the Transformer architecture can capture global dependencies, it is widely used for modality fusion (Zhang & Li, 2023; Yun et al., 2024; Jing & Zhao, 2024; Ma et al., 2024). However, these methods often overlook two critical issues: (1) inter-utterance interaction noise degrades the accuracy of attention matrices in reflecting correlation between utterances; (2) quality discrepancies among attention heads allow low-quality matrices to impair information fusion effectiveness. To address these limitations, we propose ARDT which comprises two components: Robust Head Attention and Residual Decoupled Information Fusion. ARDT serves as a core module that innovatively enhances the traditional Transformer architecture. It aims to improve the self-attention mechanism by introducing adaptive masking and residual decoupling to better handle conversational contextual noise and facilitate effective inter-modal interaction and complementarity. As all modalities share the ARDT architecture, we demonstrate it using language as the primary modality and the other two as auxiliary modalities. The whole ARDT algorithm is detailed in Algorithm 1.

Algorithm 1 Adaptive Residual Decoupled Transformer (ARDT).

Require: $\mathbf{X}_a, \mathbf{X}_v, \mathbf{X}_l$: multimodal features for acoustic, visual, and language modalities

Ensure: $\mathbf{L}, \mathbf{A}, \mathbf{V}$: enhanced multimodal representations

```

1: for each modality  $m \leftarrow a, v, l$  do
2:   for each head  $h \leftarrow 1, 2, \dots, L_h$  do
3:     Initialize learnable parameters  $\mathbf{W}_{Q_m}^h, \mathbf{W}_{K_m}^h$  for query/key projections
4:     Compute raw attention matrix  $\mathbf{o}_h^m \leftarrow (\mathbf{X}^m \mathbf{W}_{Q_m}^h)(\mathbf{X}^m \mathbf{W}_{K_m}^h)^\top$  ▷ (3)
5:     Generate implicit learnable masking matrix ▷ (4)
6:     Construct binary speaker position matrix ▷ (6)
7:     Generate the denoising attention matrix  $\tilde{\mathbf{a}}_h^m$  ▷ (5)
8:   end for
9:   Aggregate head attention matrices  $A^m \leftarrow \text{Agg}(\tilde{\mathbf{a}}_1^m, \tilde{\mathbf{a}}_2^m, \dots, \tilde{\mathbf{a}}_{L_h}^m)$ 
10:  Compute head-wise global attention scores  $\tilde{\mathbf{c}}^m \leftarrow \text{GAP}(\tilde{A}^m)$ . ▷ (7)
11:  Compute head importance weights  $\mathbf{c}^m \leftarrow \text{Sigmoid}(\text{Conv1D}(\tilde{\mathbf{c}}^m))$  ▷ (8)
12:  Refine attention matrices:  $A^m \leftarrow \mathbf{c}^m \cdot A^m$ . ▷ (9)
13:  for each head  $h \leftarrow 1, 2, \dots, L_h$  do
14:    Compute residual-aware aggregation  $\mathbf{r}_h^m$  ▷ (10)
15:  end for
16:  Concatenate  $\mathbf{r}_1^m, \mathbf{r}_2^m, \dots, \mathbf{r}_{L_h}^m$ :  $\mathbf{R}^m \leftarrow [\mathbf{r}_1^m || \mathbf{r}_2^m || \dots || \mathbf{r}_{L_h}^m]$ 
17:  Apply Layer Normalization and Feed-Forward Network to  $\mathbf{R}^m$  ▷ (11)
18: end for
19: return  $\mathbf{L}, \mathbf{A}, \mathbf{V}$ 

```

3.3.1 Robust head attention

We adopt a decoupled attention strategy (Xie et al., 2022) to enhance representations of language modality. Specifically, We first compute the attention matrix for each modality in the h -th head as follows:

$$o_h^m = (X^m W_{Q_m}^h)(X^m W_{K_m}^h)^\top, \forall h \in \{1, 2, \dots, L_h\}, \quad (3)$$

where $W_{Q_m}^h$ and $W_{K_m}^h \in \mathbb{R}^{d \times d_h}$ denote the query and key projection matrix, respectively. $O^m = \{o_1^m, o_2^m, \dots, o_{L_h}^m\} \in \mathbb{R}^{L_h \times N \times N}$ are the multi-head attention matrices for modality m . L_h indicates the number of heads and $d_h = d/L_h$.

However, the computed attention matrices O^m suffer from inter-utterance interaction noise interference, leading to unreliable correlation estimation. To solve this problem, we propose an implicit learnable masking matrices Z^m to adaptively adjust attention matrices, thereby allowing the attention matrices to better reflect the correlation between utterances:

$$Z^m = \text{Softmax}(\mu^m + \Sigma^m \odot (\lambda \cdot \epsilon)), \epsilon \sim \mathcal{N}(0, 1), \quad (4)$$

where μ^m and $\Sigma^m \in \mathbb{R}^{L_h \times N \times N}$ are learnable parameters initialized to 0 and 1, respectively. \odot denotes element-wise multiplication. A non-trainable noise scale $\lambda \in (0, 1)$ is randomly generated.

Furthermore, utterances from the same speaker exhibit inherent consistency in semantic and emotional tendencies. To leverage this prior knowledge, we construct a binary speaker position matrix $B \in \mathbb{R}^{L_h \times N \times N}$ and incorporate it into Z^m through element-wise addition:

$$\tilde{A}^m = (Z^m + B) \odot O^m, \quad (5)$$

$$B_{h,i,j} = \begin{cases} 1 & \text{if spkr}(i) = \text{spkr}(j), \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

where $\text{spkr}(i)$ indicates the speaker of the i -th utterance. The attention matrices $\tilde{A}^m = \{\tilde{a}_1^m, \tilde{a}_2^m, \dots, \tilde{a}_{L_h}^m\} \in \mathbb{R}^{L_h \times N \times N}$ are the result of combining the raw attention matrix, the implicit noise masking, and the explicit speaker-aware signals. These explicitly reflect inter-utterance speaker relationships, thereby enhancing context-aware modeling.

While traditional multi-head attention improves model performance, it neglects quality discrepancies across attention heads. Inspired by ECA-Net (Wang et al., 2020), we propose head importance learning to solve this problem. Specifically, we first compute head-wise global attention scores. Then, we compute head importance weights through inter-head interaction as follows:

$$\tilde{c}^m = \text{GAP}(\tilde{A}^m), \quad (7)$$

$$c^m = \text{Sigmoid}(\text{Conv1D}(\tilde{c}^m)), \quad (8)$$

where $\text{GAP}(\cdot)$ indicates the global average pooling and it compresses \tilde{A}^m into $\tilde{c}^m \in \mathbb{R}^{L_h}$. $\text{Conv1D}(\cdot)$ is implemented as a 1D convolutional layer with kernel size 3.

Finally, we employ importance weights $\mathbf{c}^m \in \mathbb{R}^{L_h}$ to obtain the refined attention matrices $A^m = \{a_1^m, a_2^m, \dots, a_{L_h}^m\} \in \mathbb{R}^{L_h \times N \times N}$ as follows:

$$A^m = \mathbf{c}^m \cdot \tilde{A}^m. \quad (9)$$

3.3.2 Residual decoupled information fusion

To achieve robust fusion of multimodal attention patterns while preserving primary modality integrity, we design a residual-aware aggregation mechanism as follows:

$$r_h^l = \text{Softmax}\left(\frac{o_h^l + \frac{1}{3}(a_h^a + a_h^v + a_h^l)}{\sqrt{d_h}}\right)(X^l W_{V_l}^h), \quad (10)$$

$$L = \text{LN}\left(\text{FFN}\left[r_1^l \parallel r_2^l \parallel \dots \parallel r_{L_h}^l\right]\right), \quad (11)$$

where $W_{V_l}^h$ denotes the value projection matrix. \parallel denotes the concatenation operation. $\text{FFN}(\cdot)$ and $\text{LN}(\cdot)$ are the feed-forward network and layer normalization operations, respectively. The language representation is $L = \{l_i\}_{i=1}^N \in \mathbb{R}^{N \times d}$. Similarly, the acoustic representation $A = \{a_i\}_{i=1}^N \in \mathbb{R}^{N \times d}$ and the visual representations $V = \{v_i\}_{i=1}^N \in \mathbb{R}^{N \times d}$ can be obtained.

3.4 Language-conditioned denoising learning

After obtaining enhanced representations for the three modalities, multimodal fusion is required. However, direct fusion can lead to suboptimal performance due to noise within modalities. Therefore, we propose a conditional diffusion-based denoising strategy to provide higher quality, more robust inputs for subsequent multimodal fusion. The theoretical basis is primarily built upon two key principles: 1) Addressing Modality Noise Imbalance: Prior research consistently shows that language modality outperforms acoustic and visual modalities (Tu et al., 2024; Nguyen et al., 2024; Yang et al., 2023). As illustrated in Fig. 1, real-world acoustic and visual data are often susceptible to noise, while language provides the most direct and clear emotional cues. This necessitates a targeted denoising approach for acoustic and visual signals before fusion. 2) Enhancing Denoising Capability of Diffusion Models: Diffusion model offers a powerful generative and denoising framework with a forward process to corrupt data with noise, and reverse process which iteratively recovers clean signals from the noisy input in a probabilistic manner. To enhance the capability of recovering clean signals in both acoustic and visual spaces, we resort to incorporate the high-quality language signals as condition to guide the reverse process of diffusion model for acoustic and visual representations, respectively.

To be specific, the **Language-Conditioned Denoising Learning (LDL)** module consists of three components, i.e., **Language-Conditioned Diffusion**, **Dual-Signal Alignment** and **Modality Fusion**. The whole LDL algorithm is detailed in Algorithm 2.

Algorithm 2 Language-Conditioned Denoising Learning (LDL).

Require: Enhanced multimodal representations $\mathbf{L}, \mathbf{A}, \mathbf{V}$ from ARDT
Ensure: Fused multimodal representation \mathbf{Y}

- 1: **Language-Conditioned Diffusion (Section 3.4.1):**
- 2: Initialize learnable parameters for diffusion model f_θ
- 3: **for** each modality $m \leftarrow a, v$ **do**
- 4: **Forward Diffusion Process:**
- 5: Given input \mathbf{m}_i^0 (e.g., \mathbf{A}_i or \mathbf{V}_i), inject Gaussian noise over T timesteps to get $\mathbf{m}_i^t \sim \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{m}_i^0, (1 - \bar{\alpha}_t) \mathbf{I})$ ▷ (13)
- 6: Noise schedule $1 - \bar{\alpha}_t \leftarrow s \cdot [\alpha_{\min} + \frac{t-1}{T-1}(\alpha_{\max} - \alpha_{\min})]$ ▷ (14)
- 7: **Reverse Diffusion Process:**
- 8: Reconstruct denoised features from \mathbf{m}_i^t using a language-conditioned diffusion model $f_\theta(\hat{\mathbf{m}}_i^t, \mathbf{l}_i, t)$ ▷ (16)
- 9: Optimize f_θ using loss $\mathcal{L}_d^m \leftarrow \mathbb{E}_{t, \mathbf{m}_i^0, \epsilon} [g(t) \cdot \|\mathbf{m}_i^0 - f_\theta(\hat{\mathbf{m}}_i^t, \mathbf{l}_i, t)\|^2]$ ▷ (17)
- 10: **end for**
- 11: Obtain denoised features $\hat{\mathbf{A}}$ and $\hat{\mathbf{V}}$
- 12: Total diffusion loss: $\mathcal{L}_d \leftarrow \mathcal{L}_d^a + \mathcal{L}_d^v$ ▷ (19)
- 13: **Dual-Signal Alignment (Section 3.4.2):**
- 14: **Supervised Contrastive Learning:**
- 15: **for** each modality $m \leftarrow a, v, l$ **do**
- 16: Compute supervised contrastive loss \mathcal{L}_s^m ▷ (20)
- 17: **end for**
- 18: Total supervised loss: $\mathcal{L}_s \leftarrow \mathcal{L}_s^a + \mathcal{L}_s^v + \mathcal{L}_s^l$ ▷ (21)
- 19: **Unsupervised Contrastive Learning:**
- 20: **for** each pair of modalities (m, m') **do**
- 21: Compute unsupervised contrastive loss $\mathcal{L}_u^{m, m'}$ ▷ (22)
- 22: **end for**
- 23: Total unsupervised loss: $\mathcal{L}_u \leftarrow \mathcal{L}_u^{a, v} + \mathcal{L}_u^{v, l} + \mathcal{L}_u^{a, l}$ ▷ (23)
- 24: **Modality Fusion (Section 3.4.3):**
- 25: Initialize learnable parameters for gating weights \mathbf{W}_w^a and \mathbf{W}_w^v
- 26: **for** each utterance $i \leftarrow 1, 2, \dots, N$ **do**
- 27: Compute dynamic gating weights $w_i^m \leftarrow \text{Sigmoid}([\hat{\mathbf{m}}_i \parallel \mathbf{m}_i] \mathbf{W}_w^m)$ ▷ (25)
- 28: Fuse multimodal features: $\mathbf{y}_i \leftarrow \mathbf{l}_i + \sum [(1 - w_i^m) \hat{\mathbf{m}}_i + w_i^m \mathbf{m}_i]$ ▷ (24)
- 29: **end for**
- 30: **return** $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$

3.4.1 Language-conditioned diffusion

Compared with language modality, visual and acoustic modalities usually contain more noise, which limit the ability of multimodal information fusion. Therefore, we utilize language-conditioned diffusion (Ho et al., 2020) to alleviate the noise in visual and acoustic modalities. Since the diffusion process for both visual and acoustic modality is the same, the forward and reverse processes for visual is explained in detail next.

Forward diffusion process In the forward diffusion process, we incrementally corrupt the visual features \mathbf{v}_i^0 of the i -th utterance. This is achieved by injecting Gaussian noise over T timesteps, ultimately transforming \mathbf{v}_i^0 into pure noise $\mathbf{v}_i^T \sim \mathcal{N}(0, \mathbf{I})$. Formally, given an input sample $\mathbf{v}_i^0 \sim q(\mathbf{v}_i^0)$, the Markov process generates latent variables $\{\mathbf{v}_i^0, \mathbf{v}_i^1, \dots, \mathbf{v}_i^T\}$ with the t -th step defined as:

$$q(\mathbf{v}_i^t | \mathbf{v}_i^{t-1}) = \mathcal{N}(\mathbf{v}_i^t; \sqrt{1 - \beta_t} \mathbf{v}_i^{t-1}, \beta_t \mathbf{I}), \tag{12}$$

where t runs from 1 to T , $\beta_t \in (0, 1)$ controls the noise schedule, and \mathcal{N} represents Gaussian distribution.

Using the reparameterization trick (Sohl-Dickstein et al., 2015), \mathbf{v}_t can be directly sampled from \mathbf{v}_0 via:

$$q(\mathbf{v}_i^t | \mathbf{v}_i^0) = \mathcal{N}(\mathbf{v}_i^t; \sqrt{\bar{\alpha}_t} \mathbf{v}_i^0, (1 - \bar{\alpha}_t) \mathbf{I}), \quad (13)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{t'=1}^t \alpha_{t'}$. At any timestep t , \mathbf{v}_t is sampled as $\mathbf{v}_t = \sqrt{\bar{\alpha}_t} \mathbf{v}_i^0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$.

We implement a linear noise scheduler to control the variance schedule of additive noise at each timestep (Jiang et al., 2024; Le et al., 2025):

$$1 - \bar{\alpha}_t = s \cdot \left[\alpha_{\min} + \frac{t-1}{T-1} (\alpha_{\max} - \alpha_{\min}) \right], \quad (14)$$

where $s \in [0, 1]$ controls the noise scale. α_{\min} and $\alpha_{\max} \in (0, 1)$ bound noise levels at each timestep.

Reverse diffusion process The reverse process reconstructs denoised visual features $\hat{\mathbf{v}}_i^0$ from $\hat{\mathbf{v}}_i^T$. Starting with the initial noise $\hat{\mathbf{v}}_i^T$, the denoising process as follows:

$$p_\theta(\hat{\mathbf{v}}_i^{t-1} | \hat{\mathbf{v}}_i^t) = \mathcal{N}(\hat{\mathbf{v}}_i^{t-1}; \mu_\theta(\hat{\mathbf{v}}_i^t, t), \Sigma_\theta(\hat{\mathbf{v}}_i^t, t)), \quad (15)$$

where $\hat{\mathbf{v}}_i^T = \mathbf{v}_i^T$ is the output of the forward process. The covariance $\Sigma_\theta(\hat{\mathbf{v}}_i^t, t)$ is set to $\sigma_t^2 = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$ and the mean $\mu_\theta(\hat{\mathbf{v}}_i^t, t)$ is expressed as:

$$\mu_\theta(\hat{\mathbf{v}}_i^t, t) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \hat{\mathbf{v}}_i^t + \frac{\sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)}{1 - \bar{\alpha}_t} \tilde{\mathbf{v}}_i^0, \quad (16)$$

where $\tilde{\mathbf{v}}_i^0$ is predicted by a network composed of two MLPs $f_\theta(\hat{\mathbf{v}}_i^t, \mathbf{l}_i, t)$ with the parameter θ , as \mathbf{v}_i^0 remains unknown during the reverse diffusion. Then we learn f_θ with the following loss function (Jiang et al., 2024):

$$\mathcal{L}_d^v = \mathbb{E}_{t, \mathbf{v}_i^0, \epsilon} [g(t) \cdot \|\mathbf{v}_i^0 - f_\theta(\hat{\mathbf{v}}_i^t, \mathbf{l}_i, t)\|^2], \quad (17)$$

$$g(t) = \begin{cases} \frac{\bar{\alpha}_{t-1}}{1 - \bar{\alpha}_{t-1}} - \frac{\bar{\alpha}_t}{1 - \bar{\alpha}_t}, & \text{if } t > 0, \\ 1, & \text{if } t = 0. \end{cases} \quad (18)$$

Analogously, we can obtain the acoustic modality diffusion loss \mathcal{L}_d^a through identical language-conditioned diffusion. The total diffusion loss is given by:

$$\mathcal{L}_d = \mathcal{L}_d^v + \mathcal{L}_d^a. \quad (19)$$

3.4.2 Dual-signal alignment

After obtaining denoised acoustic and visual representations $\{\hat{\mathbf{a}}_i\}_{i=1}^N$ and $\{\hat{\mathbf{v}}_i\}_{i=1}^N$, we introduce this module comprising supervised and unsupervised contrastive learning strategies:

Supervised contrastive learning For each modality m , intra-modal positive pairs are defined as utterances sharing identical emotion labels:

$$\mathcal{L}_s^m = -\frac{1}{N} \sum_{i=1}^N \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp(\hat{\mathbf{m}}_i^\top \hat{\mathbf{m}}_p / \tau)}{\sum_{k \neq i} \exp(\hat{\mathbf{m}}_i^\top \hat{\mathbf{m}}_k / \tau)}, \tag{20}$$

$$\mathcal{L}_s = \mathcal{L}_s^a + \mathcal{L}_s^v + \mathcal{L}_s^l, \tag{21}$$

where $\mathcal{P}(i)$ indexes utterances with the same label as the i -th utterance. τ is a temperature parameter, and $\hat{\mathbf{l}}_i = \mathbf{l}_i$.

Unsupervised contrastive learning For each pair of modalities m and m' , cross-modal positive pairs are constructed from same-utterance representations:

$$\mathcal{L}_u^{m,m'} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\hat{\mathbf{m}}_i^\top \hat{\mathbf{m}}_{i'} / \tau)}{\sum_{j \neq i} \exp(\hat{\mathbf{m}}_i^\top \hat{\mathbf{m}}_j / \tau)}, \tag{22}$$

$$\mathcal{L}_u = \mathcal{L}_u^{a,v} + \mathcal{L}_u^{a,l} + \mathcal{L}_u^{v,l}. \tag{23}$$

3.4.3 Modality fusion

The final representation integrates denoised and original modalities through adaptive gating weights:

$$\mathbf{y}_i = \mathbf{l}_i + \sum_{m \in \{a,v\}} [(1 - w_i^m) \hat{\mathbf{m}}_i + w_i^m \mathbf{m}_i], \tag{24}$$

$$w_i^m = \text{Sigmoid}([\hat{\mathbf{m}}_i \| \mathbf{m}_i] \mathbf{W}_w^m), \tag{25}$$

where w_i^m is a dynamic gating weight and $\mathbf{W}_w^m \in \mathbb{R}^{2d \times 1}$ is the learnable parameter. $\mathbf{y}_i \in \mathbb{R}^d$ denotes the final multimodal representation of the i -th utterance, which is fed into a classifier for emotion prediction.

3.5 Model training

3.5.1 Classifier

The emotion prediction is obtained through:

$$\mathbf{p}_i = \text{Softmax}(\mathbf{W}_c \text{ReLU}(\mathbf{y}_i) + \mathbf{b}_c), \quad (26)$$

$$\hat{e}_i = \arg \max_k \mathbf{p}_i[k], \quad (27)$$

where $\mathbf{W}_c \in \mathbb{R}^{C \times d}$ is the classification weight matrix, $\mathbf{b}_c \in \mathbb{R}^C$ the bias vector, and C the number of emotion categories. \mathbf{p}_i and \hat{e}_i represent the predicted probabilities and the predicted emotions of the i -th utterance, respectively.

3.5.2 Training Objectives

We use the cross-entropy loss as the classification loss:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^C y'_{i,k} \log \mathbf{p}_i[k], \quad (28)$$

where $y'_{i,k} \in \{0, 1\}$ is the ground-truth one-hot label of the i -th utterance for class k .

Finally, the overall training objective combines as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \gamma_s \mathcal{L}_s + \gamma_u \mathcal{L}_u + \gamma_d \mathcal{L}_d, \quad (29)$$

where γ_u , γ_s and γ_d are hyperparameters.

4 Experiments

4.1 Implementation details

Datasets We conducted experiments on two benchmark datasets: **IEMOCAP** (Busso et al., 2008) has six emotion categories (*happy, sad, neutral, angry, excited, frustrated*) and **MELD** (Poria et al., 2019) has seven categories (*neutral, surprise, fear, sadness, joy, disgust, angry*). Detailed dataset statistics are provided in Table 1.

Settings We perform all experiments on an NVIDIA GeForce RTX 4090 GPU. The software environment includes Python 3.8, Pytorch 2.4.1 and CUDA 12.1. Main hyperparameters are detailed in Table 2. Reported results are averaged over 5 runs.

Evaluation metrics We evaluate performance using Accuracy (ACC) and Weighted F1 (W-F1), with F1 scores for each emotion category.

Table 1 Statistics of IEMOCAP and MELD datasets

Dataset	Split	Conversation	Utterance	Conversation Length Statistics		
				Q1	Q3	Mean
IEMOCAP	Train+Val	120	5,810	37	59	48.42
	Test	31	1,623	42.50	59	52.35
MELD	Train+Val	1,153	11,098	5	14	9.63
	Test	280	2,610	5	13	9.32

Table 2 Main hyperparameters for DTDiff

Dataset	Parameter												
	BS	LR	Epoch	Dropout	L_h	T	s	α_{min}	α_{max}	τ	γ_u	γ_s	γ_d
IEMOCAP	18	0.001	70	0.4	5	5	0.1	0.0001	0.02	0.07	0.015	0.01	0.01
MELD	30	0.0003	70	0.15	5	5	0.1	0.0001	0.02	0.07	0.003	0.009	0.02

BS and LR represent batch size and learning rate respectively

Baselines we conduct comprehensive comparisons with ten state-of-the-art baselines: **DialogueRNN** (Majumder et al., 2019) employs recurrent networks to track speaker states; **DialogueGCN** (Ghosal et al., 2019) leverages graph convolutions for dependencies between speakers; **MMGCN** (Hu et al., 2021) fuses multimodal graphs with speaker-aware modeling; **CTNet** (Lian et al., 2021) introduces transformer-based cross-modal interactions; **MMDFN** (Hu et al., 2022) dynamically reduces modality redundancy; **SCMM** (Yang et al., 2023) adaptively selects contextual interaction paths; **CMCF-SRNet** (Zhang & Li, 2023) refines semantics via cross-modal transformers; **TopicDiff** (Luo et al., 2024) introduces a model-agnostic Topic-enriched Diffusion approach for capturing multimodal topic information; **MultiDAG** (Nguyen et al., 2024) integrates curriculum learning with directed acyclic graphs; **HAUCL** (Yi et al., 2024) combines variational hypergraphs and contrastive learning; **AdaIGN** (Tu et al., 2024) adaptively selects graph nodes via Gumbel Softmax.

4.2 Overall performances

Table 3 and 4 present the experimental results of our proposed DTDiff model on the IEMOCAP and MELD datasets, respectively. The results demonstrate that DTDiff achieves state-of-the-art overall performance on both datasets, with statistically significant improvements ($p < 0.05$) in ACC and W-F1 scores. Specifically, DTDiff improves ACC and W-F1 scores by 0.75 and 0.44 percentage points (hereafter, "points") compared to the best performing baseline AdaIGN on the IEMOCAP dataset, and by 0.68 and 0.37 points on the MELD dataset, respectively. In addition, we present a fine-grained analysis of the F1 classification scores for each emotion category. For the IEMOCAP dataset, DTDiff achieves substantial improvements for *happy* (+1.75 points) and *sad* (+1.62 points) over AdaIGN, both statistically significant ($p < 0.05$). For the MELD dataset, DTDiff shows significant gains in *neutral* (+0.29 points) and *angry* (+1.96 points).

These results confirm that DTDiff effectively captures global conversational contexts and reduces noise interference. Compared with graph-based models like AdaIGN and transformer-based methods like CMCF-SRNet, DTDiff consistently outperforms them, demonstrating its superior capability to model conversation dynamics and enhance classification accuracy for semantically complex emotions, thereby proving its robustness.

4.3 Sensitivity analysis

We conduct sensitivity analysis on three key hyperparameters in DTDiff using the IEMOCAP and MELD datasets. Figure 3 illustrates the impact of varying parameters while keeping others fixed at their optimal values.

Unsupervised contrastive learning weight (γ_u) Increasing γ_u enhances focus on dissimilarities between samples with different emotion labels, which improves emotion discrimination capability. However, excessive values (above 0.015 for IEMOCAP 0.003 for MELD) degrade performance as measured by Weighted F1-score. The relatively higher Weighted F1-score range compared with γ_s and γ_d suggests limited contribution from this component.

Supervised contrastive learning weight (γ_s) Higher γ_s emphasizes cross-modal alignment within the same utterance. Maximum Weighted F1-scores occur at $\gamma_s = 0.005$ for IEMO-

Table 3 The overall performance comparisons on IEMOCAP

Model	IEMOCAP							ACC	W-F1
	Happy	Sad	Neutral	Angry	Excited	Frustrated			
DialogueRNN [#]	32.20	80.26	57.89	62.82	73.87	59.76	63.52	62.89	
DialogueGCN [#]	51.57	80.48	57.69	53.95	72.81	57.33	63.22	62.89	
MMGCN [#]	45.14	77.16	64.36	68.82	74.71	61.40	66.36	66.26	
CTNet [‡]	51.30	79.90	65.80	67.20	78.70	58.80	68.00	67.50	
MM-DFN [‡]	42.22	78.98	66.42	69.77	75.56	66.33	68.21	68.18	
SCMM [‡]	45.37	78.76	63.54	66.05	<u>76.70</u>	66.18	–	67.53	
CMCF-SRNet [‡]	52.20	80.90	68.80	<u>70.30</u>	<u>76.70</u>	61.60	<u>70.50</u>	69.60	
TopicDif [‡]	–	–	–	–	–	–	–	67.02	
MultiDAG [‡]	49.65	81.40	69.53	70.33	71.61	66.94	69.11	69.08	
HAUCL [‡]	<u>53.57</u>	82.04	68.61	66.44	75.60	<u>68.23</u>	70.30	70.27	
AdaIGN [‡]	53.04	<u>81.47</u>	71.26	65.87	76.34	67.79	70.49	<u>70.74</u>	
DTDiff (Ours)	54.79 [†]	83.09 [†]	<u>71.01</u>	68.80	74.44	68.41	71.24 [†]	71.18 [†]	

The top performer in bold and the second best in underlined. [#] and [‡] results come from Hu et al. (2022) and original papers, respectively. [†] indicates a significant improvement over AdaIGN at $p < 0.05$

Table 4 The overall performance comparisons on MELD

Model	MELD							ACC	W-F1
	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Angry		
DialogueRNN [#]	76.97	47.69	–	20.41	50.92	–	45.52	60.31	57.66
DialogueGCN [#]	75.97	46.05	–	19.60	51.20	–	40.83	58.62	56.36
MMGCN [#]	76.33	48.15	–	26.74	53.02	–	46.09	60.42	58.31
CTNet [‡]	77.40	52.70	<u>10.00</u>	32.50	56.00	<u>11.20</u>	44.60	62.00	60.50
MM-DFN [‡]	77.76	50.69	–	22.93	54.78	–	47.82	62.49	59.46
SCMM [‡]	–	–	–	–	–	–	–	–	59.44
CMCF-SRNet [‡]	–	–	–	–	–	–	–	62.80	62.30
TopicDif [‡]	–	–	–	–	–	–	–	–	58.42
MultiDAG [‡]	–	–	–	–	–	–	–	64.41	64.00
HAUCL [‡]	–	–	–	–	–	–	–	<u>68.05</u>	66.72
AdaIGN [‡]	<u>79.75</u>	60.53	–	<u>43.70</u>	64.54	–	<u>56.15</u>	67.62	<u>66.79</u>
DTDiff (Ours)	80.04 [†]	<u>58.97</u>	28.41	43.79	<u>64.35</u>	26.64	57.91 [†]	68.30 [†]	67.16 [†]

The top performer in bold and the second best in underlined. [#] and [‡] results come from Hu et al. (2022) and original papers, respectively. [†] indicates a significant improvement over AdaIGN at $p < 0.05$

CAP and $\gamma_s = 0.009$ for MELD. The significant performance drop at $\gamma_s = 0$ confirms the importance of intra-utterance multimodal alignment.

Diffusion loss weight (γ_d) This parameter balances denoised and original feature learning. Optimal performance is achieved at $\gamma_d = 0.01$ for IEMOCAP and $\gamma_d = 0.02$ for MELD. Both insufficient and excessive weighting degrade model effectiveness, demonstrating the necessity of balanced denoising supervision.

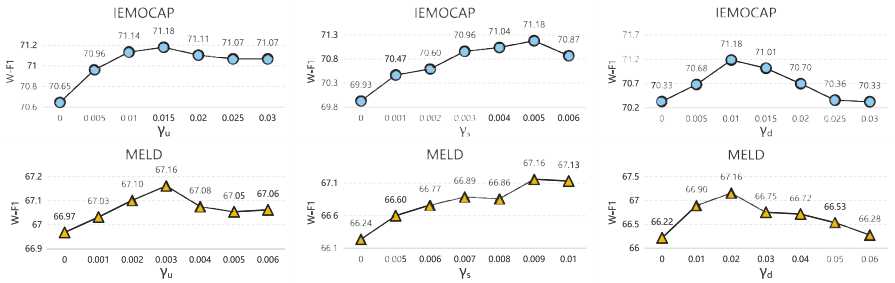


Fig. 3 Sensitive analysis of DTDiff on IEMOCAP and MELD. All experiments test the results while fixing all other parameters with the best performance.

Table 5 Ablation studies for different modalities, where L, A, and V denote Language, Acoustic, and Visual, respectively

	IEMOCAP		MELD	
	ACC	W-F1	ACC	W-F1
Only A	58.87	58.21	48.34	43.35
Only V	45.24	42.50	47.47	33.56
Only L	68.05	68.15	66.93	66.09
A+V	61.88	61.03	49.96	45.24
V+L	69.24	69.09	67.04	66.21
A+L	69.84	69.74	67.23	66.41
A+V+L (Ours)	71.24	71.18	68.30	67.16

4.4 Ablation study

4.4.1 Ablation studies for modalities

The experiments presented in Table 5 investigate the impact of individual modalities and their combinations on model performance. The results clearly demonstrate that the performance of the model using multimodal data is superior to that using single-modality data, and the performance of the language modality is far better than that of the other two modalities. This further confirms that the acoustic and visual modalities have more noise than the language modality. In addition, a comparison of L and A+V+L reveals that the model performs better when incorporating the information from the other two modalities compared to using only the language modality. This can be attributed to the fact that by integrating information from multiple modalities, a more comprehensive and detailed representation can be obtained, thereby enabling the model to capture complex emotional cues in conversations.

4.4.2 Ablation studies for ARDT and LDL

To investigate the contributions of each component in DTDiff, we conduct ablation studies on both datasets. The results are shown in Table 6 and the experimental findings lead to the following conclusions: (1) Removing the entire AMDT (**w/o AMDT**) leads to a significant performance drop, demonstrating that the transformer-based architecture enhances the representation of each modality and improves model robustness. (2) Disabling the multi-head attention weights in AMDT (**w/o WAH**) reduces performance, indicating that Weighting

Table 6 Performance of DTDiff for ablation study

Methods	IEMOCAP		MELD	
	ACC	W-F1	ACC	W-F1
w/o ARDT	69.60	69.54	67.36	66.34
w/o WAH	70.72	70.29	67.44	66.43
w/o LDL	69.76	69.56	67.42	66.17
w/o condition L	70.32	69.93	67.65	66.18
w/o \hat{A} and \hat{V}	70.94	70.51	67.99	66.69
w/ condition A	70.78	70.25	66.70	65.34
w/ condition V	70.66	70.65	66.74	65.37
Ours	71.24	71.18	68.30	67.16

Table 7 Ablation study on different masking signal (B and Z) combinations

B	Z	IEMOCAP		MELD	
		ACC	W-F1	ACC	W-F1
✓	✓	71.24	71.18	68.30	67.16
✓	✗	70.66	70.30	67.66	66.42
✗	✓	70.38	70.12	67.53	66.46
✗	✗	70.09	70.06	67.43	66.40

Attention Heads via inter-matrix interactions effectively identifies high-quality attention matrices. Compared to removing the entire AMDT, retaining WAH with adaptive masking operations significantly mitigates performance degradation, proving that dynamic masking critically filters inter-utterance interaction noise while preserving utterance relevance. (3) Removing the entire LDL (**w/o LDL**) severely degrades performance, validating the necessity of language constraints in suppressing modality-specific noise from acoustic and visual inputs. (4) Replacing language-conditioned diffusion with standard diffusion (**w/o condition L**) shows that the cues in language modality can denoise the acoustic and visual modalities. While standard diffusion fails to denoise effectively, the Dual-Signal Alignment module partially compensates by aligning modality features. (5) Excluding denoised audio and visual features (**w/o \hat{A} and \hat{V}**) during fusion causes performance decline, confirming that LDL not only reduces modality-specific noise but also learns denoising features weights for optimal fusion. (6) Replacing language-conditioned diffusion with acoustic-conditioned diffusion (**w/ condition A**) or visual-conditioned diffusion (**w/ condition V**) results in performance degradation, especially on MELD, confirming that language modality provides the most effective conditioning for denoising due to its clearer emotional cues and less noise.

In conclusion, AMDT and LDL collaboratively enhance performance. Specifically, AMDT resolves inter-utterance interaction noise through adaptive masking and representation enhancement, while LDL tackles modality-specific noise via language-conditioned denoising and alignment. Their synergy enables robust multimodal emotion understanding.

4.4.3 Ablation studies for masking signals

As demonstrated in Table 7, we investigate the impact of two masking signals. The results show that removing both B and Z leads to the most significant performance degradation, confirming the synergistic effect of prior and posterior knowledge in noise suppression.

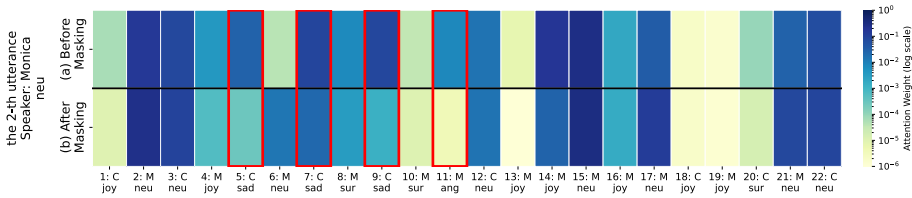


Fig. 4 Visualization of attention weights between the second utterance and other utterances in a MELD conversation, before (a) and after (b) applying masking signals. Speakers are denoted by C (Chandler) and M (Monica). The true emotion labels for each utterance are indicated below the speaker ID: 'joy' (joy), 'neu' (neutral), 'sad' (sadness), 'sur' (surprise), and 'ang' (angry)

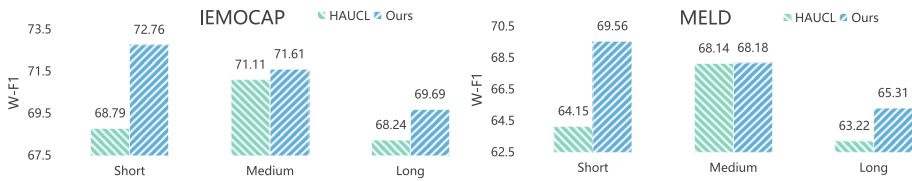


Fig. 5 Comparison of the performance of HAUCL and DTDiff under different conversation lengths on IEMOCAP and MELD

Furthermore, introducing either Z or B yields better performance than removing both, proving that both masking signals are effective for DTDiff and that speaker position information alone is insufficient for comprehensive contextual modeling and requires integration with noise suppression mechanisms.

Figure 4 visualizes the attention scores between the second utterance and 22 others in a conversation between speakers Chandler (C) and Monica (M) from the MELD test set. The first row displays the attention scores before masking, while the second row shows results after masking. The comparison reveals that the attention correlation between the second utterance (*neutral*) and negative-emotion utterances like the fifth, seventh, ninth (*sad*) and eleventh (*angry*) utterances is significantly reduced. Since the conversational context is dominated by neutral or positive emotions, the masking operation effectively suppresses interference from irrelevant negative emotions through explicit positional signals and implicit denoising signals.

4.5 Analysis of conversation length

As shown in Table 1, we adopt quantiles to group conversations in the test set based on conversation length distributions of the IEMOCAP and MELD datasets. Specifically for IEMOCAP: short conversations (length ≤ 43), medium conversations ($43 < \text{length} \leq 59$), and long conversations (length > 59). The MELD dataset exhibits significantly shorter conversations with short conversation (length ≤ 5), medium conversations ($5 < \text{length} \leq 13$), and long conversations (length > 13).

Figure 5 illustrates our proposed DTDiff model's performance across various conversation lengths compared to HAUCL. Crucially, DTDiff consistently outperforms HAUCL in all three conversation length categories. This demonstrates DTDiff's enhanced capability in handling conversations of varying durations more effectively than HAUCL. Specifically,

DTDiff's transformer-based architecture, with its adaptive masking mechanism, effectively filters irrelevant contextual dependencies and mitigates noise. It offers a significant advantage over graph-based methods like HAUCL which suffers from excessive redundancy in contextual messages and over-smoothing in graph network, particularly in long conversations with fully connected graph structures. Moreover, we observe a general trend where DTDiff's performance, while still superior to HAUCL, gradually degrades as conversation length increases. This phenomenon is primarily attributed to the significantly increased complexity of contextual dependencies in longer conversations. As conversation length grows, semantic and emotional cues become more intricate. Although DTDiff employs dynamic masking and denoising to filter noise, it still encounters challenges in capturing all critical, long-range contextual information. To address this limitation, we will study enhancing long-conversation comprehension through hierarchical contextual modeling, or optimizing long-distance dependencies via more efficient attention mechanisms and auxiliary tasks in future work.

5 Limitations

Despite extensive experimental validation demonstrating the effectiveness of Transformer-based architecture and language-conditioned diffusion model within our DTDiff framework for MERC, certain limitations is worth of consideration. Primarily, when emotional cues in the language modality are ambiguous or weak, the denoising efficacy for the acoustic and visual modalities can be compromised. In such scenarios, the conditioning signal from language might be insufficient or even misleading, leading to suboptimal denoising outcomes and subsequently impacting the overall model performance. Furthermore, given that DTDiff incorporates a Transformer architecture and conditional diffusion model, it inherently introduces significant computational cost, especially when dealing with longer conversations (i.e., large N values). These limitations highlight promising avenues for future research, including exploring more robust cross-modal condition mechanisms that can tolerate weaker linguistic signals. Additionally, optimizing transformer's attention strategy and the sampling steps of diffusion model could enhance efficiency and scalability for more practical applications.

6 Conclusion

In this paper, we propose **DTDiff**, a multimodal emotion recognition model for conversational scenarios that integrates a dynamic masking mechanism and language-conditioned diffusion to systematically address dual noise challenges: cross-utterance interaction noise and modality-specific noise. By adaptively suppressing low-quality attention heads and enhancing feature compatibility through Dual-Signal Alignment, the model achieves significant improvements in robustness and accuracy. Experimental results demonstrate the superior performance of DTDiff over existing methods. Future research could focus on enhancing long-conversation comprehension, resolving ambiguous emotion distinctions, and mitigating class imbalance issues to further advance practical applications.

Author Contributions T.T. conducted experiments and prepared the original draft. T.T. and X.Z. supervised the project and contributed to reviewing, editing, and writing. All authors reviewed the manuscript.

Funding This work was supported by the Science and Technology Innovation Key R&D Program of Chongqing under Grant CSTB2024TIAD-STX0027 the National Natural Science Foundation of China under Grant 62472059 the Chongqing Talent Plan Project, China under Grant CSTC2024YCYH-BGZX0022.

Data Availability The datasets used in this study are publicly accessible. The IEMOCAP dataset (<https://paperswithcode.com/dataset/iemocap>) and the MELD dataset (<https://affective-meld.github.io/>).

Declarations

Competing interests The authors declare no competing interests.

Ethical approval Not Applicable.

References

- Busso, C., Bulut, M., Lee, C., et al. (2008). Iemocap: interactive emotional dyadic motion capture database. *Lang Resour Evaluation*, 42(4), 335–359. <https://doi.org/10.1007/S10579-008-9076-6>
- Chen, F., Shao, J., Zhu, S., et al. (2023). Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In: IEEE/CVF conference on computer vision and pattern recognition, pp 10761–10770. <https://doi.org/10.18653/V1/D18-1280>
- Chen, B., Zhang, Z., Li, W., et al. (2024). Invertible diffusion models for compressed sensing. CoRR [arXiv:2403.17006](https://arxiv.org/abs/2403.17006). <https://doi.org/10.1109/TPAMI.2025.3538896>
- Gan, X., Huang, X., & Zou, S. (2025). Intentional tendency-based dynamic heterogeneous graph network for emotion recognition in conversations. *Journal of Intelligent Information Systems*, 63(3), 989–1010. <https://doi.org/10.1007/S10844-025-00925-9>
- Ghosal, D., Majumder, N., Poria, S., et al. (2019). Dialoguegen: A graph convolutional neural network for emotion recognition in conversation. In: Proceedings of the 2019 conference on empirical methods in natural language processing, pp. 154–164, 10.18653/V1/D19-1015
- Ho, J., Jain, A., Abbeel, P. (2020). Denoising diffusion probabilistic models. In: Advances in neural information processing systems 33: annual conference on neural information processing systems 2020. <https://doi.org/10.48550/arXiv.2006.11239>
- Hu, D., Hou, X., Wei, L., et al. (2022). MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. In: IEEE international conference on acoustics, speech and signal processing, pp. 7037–7041. <https://doi.org/10.1109/ICASSP43922.2022.9747397>
- Hu, J., Liu, Y., Zhao, J., et al. (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, pp. 5666–5675. <https://doi.org/10.18653/V1/2021.ACL-LONG.440>
- Ishiwatari, T., Yasuda, Y., Miyazaki, T., et al. (2020). Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In: Proceedings of the 2020 conference on empirical methods in natural language processing, pp. 7360–7370. <https://doi.org/10.18653/V1/2020.EMNLP-MAIN.597>
- Jannu, C., Vanambathina, S.D. (2023a). An attention based densely connected u-net with convolutional gru for speech enhancement. In: 2023 3rd International conference on artificial intelligence and signal processing (AISP), pp. 1–5. <https://doi.org/10.1109/AISP57993.2023.10134933>
- Jannu, C., Vanambathina, S.D. (2023b). Convolutional transformer based local and global feature learning for speech enhancement. *International Journal of Advanced Computer Science and Applications* 14(1). 10.14569/IJACSA.2023.0140181
- Jannu, C., & Vanambathina, S. D. (2023). Dct based densely connected convolutional gru for real-time speech enhancement. *Journal of Intelligent & Fuzzy Systems*, 45(1), 1195–1208. <https://doi.org/10.3233/JIFS-223951>

- Jannu, C., & Vanambathina, S. D. (2025). An overview of speech enhancement based on deep learning techniques. *International Journal of Image and Graphics*, 25(01), 2550001. <https://doi.org/10.1142/S0219467825500019>
- Jiang, Y., Xia, L., Wei, W., et al. (2024). Diffimm: Multi-modal diffusion model for recommendation. In: Proceedings of the 32nd ACM international conference on multimedia, pp. 7591–7599. <https://doi.org/10.1145/3664647.3681498>
- Jiao, W., Yang, H., King, L., et al. (2019). Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, pp. 397–406. <https://doi.org/10.18653/V1/N19-1037>
- Jing, Y., Zhao, X. (2024). Dq-former: Querying transformer with dynamic modality priority for cognitive-aligned multimodal emotion recognition in conversation. In: Proceedings of the 32nd ACM international conference on multimedia, pp 4795–4804. <https://doi.org/10.1145/3664647.3681599>
- Job, S., Tao, X., Cai, T., et al. (2025). Exploring causal learning through graph neural networks: an in-depth review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15. <https://doi.org/10.1002/widm.70024>
- Joshi, A., Bhat, A., Jain, A., et al. (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, pp. 4148–4164. <https://doi.org/10.18653/v1/2022.naacl-main.306>
- Le, Y., Li, H., Ou, B., et al. (2025). Diffusion model for interest refinement in multi-interest recommendation. <https://doi.org/10.48550/ARXIV.2502.05561>
- Li, B., Fei, H., Liao, L., et al. (2023a). Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In: Proceedings of the 31st ACM international conference on multimedia, pp. 5923–5934. <https://doi.org/10.1145/3581783.3612053>
- Li, J., Ji, D., Li, F., et al. (2020). Hitrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In: Proceedings of the 28th international conference on computational linguistics, pp. 4190–4200. <https://doi.org/10.18653/V1/2020.COLING-MAIN.370>
- Li, X.L., Thickstun, J., Gulrajani, I., et al. (2022). Diffusion-lm improves controllable text generation. In: Advances in neural information processing systems 35: annual conference on neural information processing systems 2022. <https://doi.org/10.48550/arXiv.2205.14217>
- Li, Y., Wang, Y., Cui, Z. (2023c). Decoupled multimodal distilling for emotion recognition. In: Proceedings of the 2023 IEEE/CVF conference on computer vision and pattern recognition, pp. 6631–6640. <https://doi.org/10.1109/CVPR52729.2023.00641>
- Li, D., Wang, Y., Funakoshi, K., et al. (2023b). Joyful: Joint modality fusion and graph contrastive learning for multimodal emotion recognition. <https://doi.org/10.48550/ARXIV.2311.11009>
- Li, Z., Xia, L., Huang, C. (2024b). Recdiff: Diffusion model for social recommendation. In: Proceedings of the 33rd ACM international conference on information and knowledge management, pp. 1346–1355. <https://doi.org/10.1145/3627673.3679630>
- Lian, Z., Liu, B., & Tao, J. (2021). Ctnet: conversational transformer network for emotion recognition. *IEEE ACM Trans Audio Speech Lang Process*, 29, 985–1000. <https://doi.org/10.1109/TASLP.2021.3049898>
- Li, G., Jin, D., Zheng, Y., et al. (2024). A generic plug & play diffusion-based denosing module for medical image segmentation. *Neural Networks*, 172, Article 106096. <https://doi.org/10.1016/j.neunet.2024.106096>
- Lin, Y., Cheng, H., Huang, C., et al. (2025). Impact of glyph information on latent space diffusion models for accurate handwritten text generation. In: 2025 IEEE international conference on acoustics, speech and signal processing, pp. 1–5. <https://doi.org/10.1109/ICASSP49660.2025.10890644>
- Luo, J., Wang, J., Zhou, G. (2024). 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, pp. 16304–16314. <https://doi.org/10.48550/arXiv.2403.04789>
- Majumder, N., Poria, S., Hazarika, D., et al. (2019). Dialoguernn: An attentive RNN for emotion detection in conversations. In: Proceedings of the 33rd AAAI conference on artificial intelligence, AAAI 2019, pp. 6818–6825. <https://doi.org/10.1609/AAAI.V33I01.33016818>
- Ma, H., Wang, J., Lin, H., et al. (2024). A transformer-based model with self-distillation for multimodal emotion recognition in conversations. *IEEE Trans Multim*, 26, 776–788. <https://doi.org/10.1109/TMM.2023.3271019>
- Nguyen, C.T., Nguyen, C., Le, D., et al. (2024). Curriculum learning meets directed acyclic graph for multimodal emotion recognition. In: Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation, pp. 4259–4265. <https://doi.org/10.48550/arXiv.2402.17269>

- Parisae, V., & Nagakishore Bhavanam, S. (2024). Multi scale encoder-decoder network with time frequency attention and s-tn for single channel speech enhancement. *Journal of Intelligent & Fuzzy Systems*, 46(4), 10907–10907. <https://doi.org/10.3233/JIFS-233312>
- Poria, S., Cambria, E., Hazarika, D., et al. (2017). Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th annual meeting of the association for computational linguistics, pp. 873–883. <https://doi.org/10.18653/V1/P17-1081>
- Poria, S., Hazarika, D., Majumder, N., et al. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In: Proceedings of the 57th conference of the association for computational linguistics, pp. 527–536. <https://doi.org/10.18653/V1/P19-1050>
- Shen, W., Wu, S., Yang, Y., et al. (2021). Directed acyclic graph network for conversational emotion recognition. In: Proceedings of the 59th annual meeting of the association for computational linguistics, pp. 1551–1560. <https://doi.org/10.18653/V1/2021.ACL-LONG.123>
- Shi, Y., Cai, J., & Liao, L. (2025). Multi-task learning and mutual information maximization with crossmodal transformer for multimodal sentiment analysis. *Journal of Intelligent Information Systems*, 63(1), 1–19. <https://doi.org/10.1007/S10844-024-00858-9>
- Shou, Y., Ai, W., Du, J., et al. (2024). Efficient long-distance latent relation-aware graph neural network for multi-modal emotion recognition in conversations. <https://doi.org/10.48550/ARXIV.2407.00119>
- Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., et al. (2015). Deep unsupervised learning using non-equilibrium thermodynamics. In: Proceedings of the 32nd international conference on machine learning, pp. 2256–2265. <https://doi.org/10.48550/arXiv.1503.03585>
- Sun, J., Han, S., Ruan, Y., et al. (2023). Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In: Proceedings of the 61st annual meeting of the association for computational linguistics, pp. 658–670. <https://doi.org/10.18653/V1/2023.ACL-LONG.39>
- Sun, K., Chen, Z., Lin, X., et al. (2025). Conditional diffusion models for camouflaged and salient object detection. *IEEE Trans Pattern Anal Mach Intell*, 47(4), 2833–2848. <https://doi.org/10.1109/TPAMI.2025.3527469>
- Suryanto, N., Adiputra, A.A., Kadiptya, A.Y., et al. (2025) Cityscape-adverse: Benchmarking robustness of semantic segmentation with realistic scene modifications via diffusion-based image editing. pp 69921–69940. <https://doi.org/10.1109/ACCESS.2025.3537981>
- Tang, D., Cao, X., Hou, X., et al. (2024). Crs-diff: controllable remote sensing image generation with diffusion model. *IEEE Trans Geosci Remote Sens*, 62, 1–14. <https://doi.org/10.1109/TGRS.2024.3453414>
- Tu, G., Xie, T., Liang, B., et al. (2024). Adaptive graph learning for multimodal conversational emotion detection. In: Proceedings of the 38th AAAI conference on artificial intelligence, pp. 19089–19097. <https://doi.org/10.1609/AAAI.V38I17.29876>
- Wang, Q., Wu, B., Zhu, P., et al. (2020) Eca-net: Efficient channel attention for deep convolutional neural networks. In: 2020 IEEE/CVF conference on computer vision and pattern recognition, pp. 11531–11539. <https://doi.org/10.1109/CVPR42600.2020.01155>
- Wu, J., Liu, J., Zhang, T., et al. (2025). a^2h^2 for multimodal emotional data analysis. *Journal of Intelligent Information Systems*. <https://doi.org/10.1007/s10844-025-00974-0>
- Wu, Z., Zhang, Q., Miao, D., et al. (2024). Hydiscgan: A hybrid distributed cgan for audio-visual privacy preservation in multimodal sentiment analysis. In: Proceedings of the 33rd international joint conference on artificial intelligence, pp. 6550–6558. <https://doi.org/10.24963/ijcai.2024/724>
- Wu, J., Wu, J., Zheng, Y., et al. (2025). Mlgat: multi-layer graph attention networks for multimodal emotion recognition in conversations. *Journal of Intelligent Information Systems*, 63(2), 375–394. <https://doi.org/10.1007/S10844-024-00879-4>
- Xie, Y., Zhou, P., Kim, S. (2022). Decoupled side information fusion for sequential recommendation. In: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, pp. 1611–1621. <https://doi.org/10.1145/3477495.3531963>
- Yang, H., Gao, X., Wu, J., et al. (2023). Self-adaptive context and modal-interaction modeling for multimodal emotion recognition. In: Findings of the association for computational linguistics, pp. 6267–6281. <https://doi.org/10.18653/V1/2023.FINDINGS-ACL.390>
- Yi Z, Zhao Z, Shen Z, et al (2024) Multimodal fusion via hypergraph autoencoder and contrastive learning for emotion recognition in conversation. In: Proceedings of the 32nd ACM international conference on multimedia4, pp. 4341–4348. <https://doi.org/10.1145/3664647.3681633>
- Yue, Y., Yu, M., Yang, L., et al. (2025). Joint conditional diffusion model for image restoration with mixed degradations. *Neurocomputing*, 626, Article 129512. <https://doi.org/10.1016/J.NEUCOM.2025.129512>
- Yun, T., Lim, H., Lee, J., et al. (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, pp. 82–95. <https://doi.org/10.18653/V1/2024.NAAACL-LONG.5>

- Zhang, X., Li, Y. (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, pp. 13099–13110. <https://doi.org/10.18653/V1/2023.ACL-LONG.732>
- Zhang, Y., Long, J., & Li, C. (2025). Knowledge distillation for object detection with diffusion model. *Neurocomputing*, 636, Article 130019. <https://doi.org/10.1016/J.NEUCOM.2025.130019>
- Zhang, H., Xu, H., Long, F., et al. (2024a) Unsupervised multimodal clustering for semantics discovery in multimodal utterances. In: Proceedings of the 62nd annual meeting of the association for computational linguistics, pp. 18–35. <https://doi.org/10.18653/V1/2024.ACL-LONG.2>
- Zhang, M., Cai, Z., Pan, L., et al. (2024). Motiondiffuse: text-driven human motion generation with diffusion model. *IEEE Trans Pattern Anal Mach Intell*, 46(6), 4115–4128. <https://doi.org/10.1109/TPAMI.2024.3355414>
- Zheng, X., Zhao, G., Zhu, L., et al. (2022). PERD: personalized emoji recommendation with dynamic user preference. In: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, pp. 1922–1926. <https://doi.org/10.1145/3477495.3531779>
- Zong, D., Ding, C., Li, B., et al. (2023). Acformer: An aligned and compact transformer for multimodal sentiment analysis. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 833–842. <https://doi.org/10.1145/3581783.3611974>
- Zou, S., Huang, X., Shen, X. (2023). Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation. In: Proceedings of the 31st ACM international conference on multimedia, pp. 5994–6003. <https://doi.org/10.1145/3581783.3611805>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.