DELTA: Disentangled Hierarchical Interaction and Adaptive Adjustment for Emotion and Intent Understanding in Multimodal Conversations

Shenjie Jiang[1] , **Xiangfeng Liu**[1] , **Xianghua Li**[1,*] , **Xuecheng Wu**[2]
[1]Northwestern Polytechnical University, Xi'an, China
[2]Xi'an Jiaotong University, Xi'an, China
{jiangshenjie, xiangfengliu}@mail.nwpu.edu.cn, li_xianghua@nwpu.edu.cn, xuecwu@gmail.com

Abstract

Emotion and intent joint understanding in multimodal conversations (MC-EIU) aims to infer emotion and intent information by leveraging the semantic dependencies among multimodal data. However, prior works largely ignore redundancy interference and suffer from insufficient interaction and inter-task noise propagation due to their reliance on shallow mechanisms. To overcome these limitations, we propose a novel framework named **D**isentangled Hi**e**rarchica**l** Int**e**raction and Adaptive **A**djustment (DELTA) for MC-EIU. We first design a Disentangled Feature Denoising module based on orthogonal decomposition to effectively filter out noise and redundancy from heterogeneous data. Second, we propose a Dual-Level Adaptive Adjustment mechanism to dynamically optimize learning dynamics from both modality and instance perspectives. Furthermore, we introduce a Hierarchical Task-Bottleneck Interaction(HTBI) module, which employs a set of progressively compressed bottleneck tokens as a communication hub. This design can effectively simulating the multi-round iterative interaction between emotion and intent tasks. Experiments on the benchmark MC-EIU bilingual dataset demonstrate that our framework significantly outperforms state-of-the-art baselines in both emotion and intent tasks.

1 Introduction

Emotion and intent joint understanding in multimodal conversations [Singh *et al.*, 2022; Liu *et al.*, 2024; Liu *et al.*, 2026] is an emerging task that aims to jointly infer emotional states and behavioral intents from multimodal inputs, including audio, video, and text content. Unlike traditional methods that treat emotion and intent recognition as independent problems, MC-EIU emphasizes the semantic interaction between them, leveraging their interdependence to enhance joint understanding. This task holds immense potential for broad applications, such as dialogue systems [Danieli *et al.*, 2015], recommendation systems [Li *et al.*, 2024], mental health assessment tools [Francese and Attanasio, 2023], social robots [Wu *et al.*, 2025], and customer service agents [Gong *et al.*, 2022].

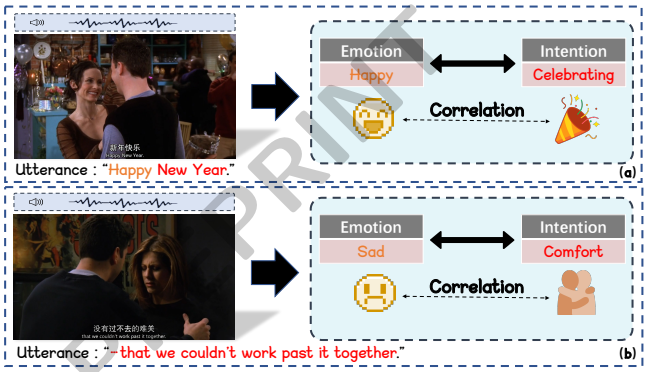

Figure 1: Motivation of our proposed DELTA framework. Emotion and intent share a close relationship that can be leveraged to enhance joint understanding. **(i)** Emotion provides key cues for intent recognition. For example, in case (a), the *Celebrating* intent serves as a strong signal to help identify the *Happy* emotion. **(ii)** The interaction is mutually beneficial. As shown in case (b), the *Sad* emotion helps infer the *Comforting* intent, demonstrating the two tasks can facilitate each other.

In recent years, the research focus has transitioned from independent emotion recognition [Li *et al.*, 2023c; Xia *et al.*, 2025] or intent recognition [Yang *et al.*, 2024] towards the joint MC-EIU task. This shift has led to the development of several representative frameworks. For instance, Singh et al. [Singh *et al.*, 2022] proposed EmoInt-Trans, a multimodal contextual Transformer for joint recognition. Liu et al. [Liu *et al.*, 2024] introduced the EI^2 network to model the correlation between emotion and intent, while Zheng et al. [Zheng *et al.*, 2025] developed MEI-Pro to leverage progressive interactions. While these methods have achieved promising performance, certain limitations remain. Regarding multimodal processing, current models often lack explicit mechanisms to decouple redundant noise from essential semantics or to dynamically distinguish the varying contributions of distinct modalities. Consequently, irrelevant signals may interfere with feature learning. Moreover, concerning task synergy, most approaches primarily rely on shallow, single-round interaction mechanisms. Such designs may not fully capture the deep, iterative mutual reinforcement between emotion and

*Corresponding author.

intent, and risk propagating noise from one task to the other.

A primary challenge lies in the interference from multimodal noise and the neglect of dynamic learning variations [Zhu *et al.*, 2019; Li *et al.*, 2023c]. Multimodal data naturally contains noise and redundancy. These interfering signals degrade the quality of feature representations, making it difficult to extract consistent emotion and intent semantics. Moreover, existing methods often overlook that the contribution of each modality varies dynamically. Relying on a static fusion strategy may allow a dominant but irrelevant modality to overshadow critical subtle cues provided by others. Additionally, the training process is often dominated by simple samples, limiting the model's ability to generalize to hard samples with complex multimodal conflicts. Therefore, to achieve robust understanding, it is essential to effectively filter out noise and adaptively adjust the learning focus across both modalities and instances.

Another key challenge lies in modeling the deep and iterative interaction between emotion and intent [Singh *et al.*, 2022]. In real-world dialogue, emotion and intent are not independent. As shown in Figure 1, they mutually influence each other. However, current approaches primarily employ single-round direct interactions, which may not fully capture this dynamic interplay. Worse still, these direct pathways risk becoming "noise corridors," where ambiguous signals or errors from one task could potentially propagate to the other. This lead to error accumulation instead of the desired mutual enhancement. To address these challenges and achieve more effective joint understanding, two fundamental questions need to be carefully considered: **Q1: How can we effectively decouple redundancy and noise from essential semantics, and dynamically balance both modality contributions and the learning focus on samples of varying difficulty? Q2: How can we achieve robust interaction and mutual enhancement between emotion and intent while simultaneously filtering out potential noise?**

To overcome the aforementioned limitations, we propose DELTA, a novel Disentangled Hierarchical Interaction and Adaptive Adjustment framework for MC-EIU. Firstly, to eliminate redundancy and noise, we introduce a Disentangled Feature Denoising (DFD) module. By leveraging orthogonal decomposition, we project raw inputs into modality-invariant and modality-specific subspaces. This enables us to effectively filter out inherent redundancy and conflicting noise while retaining the essential information shared across modalities. Secondly, we devise a Dual-Level Adaptive Adjustment mechanism to calibrate learning dynamics. This strategy dynamically re-weights modalities based on their contextual relevance and modulates instance-level gradients. By doing so, we can effectively distinguish the varying contributions of different modalities and rectify the imbalance between simple and hard samples, preventing the model from being dominated by easy patterns. Finally, we replace shallow fusion with a Hierarchical Task-Bottleneck Interaction (HTBI) module. Acting as an iterative communication hub with progressive compression, HTBI facilitates a multi-round, noise-filtered bidirectional flow of information. This process mimics the dynamic, step-by-step interplay between emotion and intent, allowing for deep interaction and mutual optimization while preventing noise propagation.

The main contributions can be summarized as follows:

- We propose a Disentangled Feature Denoising (DFD) module combined with a Dual-Level Adaptive Adjustment mechanism, which effectively purifies multimodal representations, dynamically balances contributions across heterogeneous modalities, and adapts the optimization focus based on sample difficulty.
- We develop the Hierarchical Task-Bottleneck Interaction (HTBI) strategy, a progressive interaction paradigm that captures the deep-seated, iterative dependencies between emotion and intent, preventing noise propagation while fostering mutual reinforcement.
- Extensive experiments demonstrate that DELTA establishes new state-of-the-art results on standard benchmarks, proving the efficacy of our hierarchical and adaptive design.

2 Related Work

2.1 Disentangled Multimodal Representation Learning

Disentangled multimodal representation learning aims to decompose complex multimodal data into independent latent factors, typically separating modality-invariant and modality-specific components [Tsai *et al.*, 2018; Jiang *et al.*, 2025]. [Hazarika *et al.*, 2020] proposed MISA, which projects heterogeneous modalities into common and private subspaces using orthogonality constraints to reduce disparities. Building on this, recent advancements have incorporated adversarial training [Yang *et al.*, 2022a] or distillation techniques [Li *et al.*, 2023c] to further enhance feature independence and fusion efficacy. [Li *et al.*, 2023a] propose a dual-level disentanglement mechanism and adaptive fusion strategies to enhance multimodal emotion analysis. [Fang *et al.*, 2025] propose EMOE, which utilizes a mixture of modality experts and unimodal distillation to dynamically adjust modality importance and retain unimodal predictive capabilities. However, most existing methods primarily utilize disentanglement for feature alignment or general fusion, often overlooking the critical issue of inherent redundancy and conflicting noise within the data. In our DELTA framework, the primary goal of disentanglement is feature denoising. We employ an orthogonal decomposition strategy with explicit constraints to rigorously filter out redundancy and conflicting signals, ensuring that only purified, complementary semantics are retained to facilitate accurate joint emotion and intent understanding.

2.2 Emotion and Intent Joint Understanding in Multimodal Conversations

The task of Unified Multimodal Emotion-Intent Emotion and intent joint understanding in multimodal conversations (MC-EIU) aims to simultaneously predict emotional states and behavioral intents from heterogeneous data, seeking a comprehensive grasp of user expressions. Recently, several studies have been dedicated to developing unified frameworks for this purpose. For instance, [Singh *et al.*, 2022] proposed the EmoInt-MD dataset and proposed a multi-task transformer framework that utilizes shared contextual embeddings

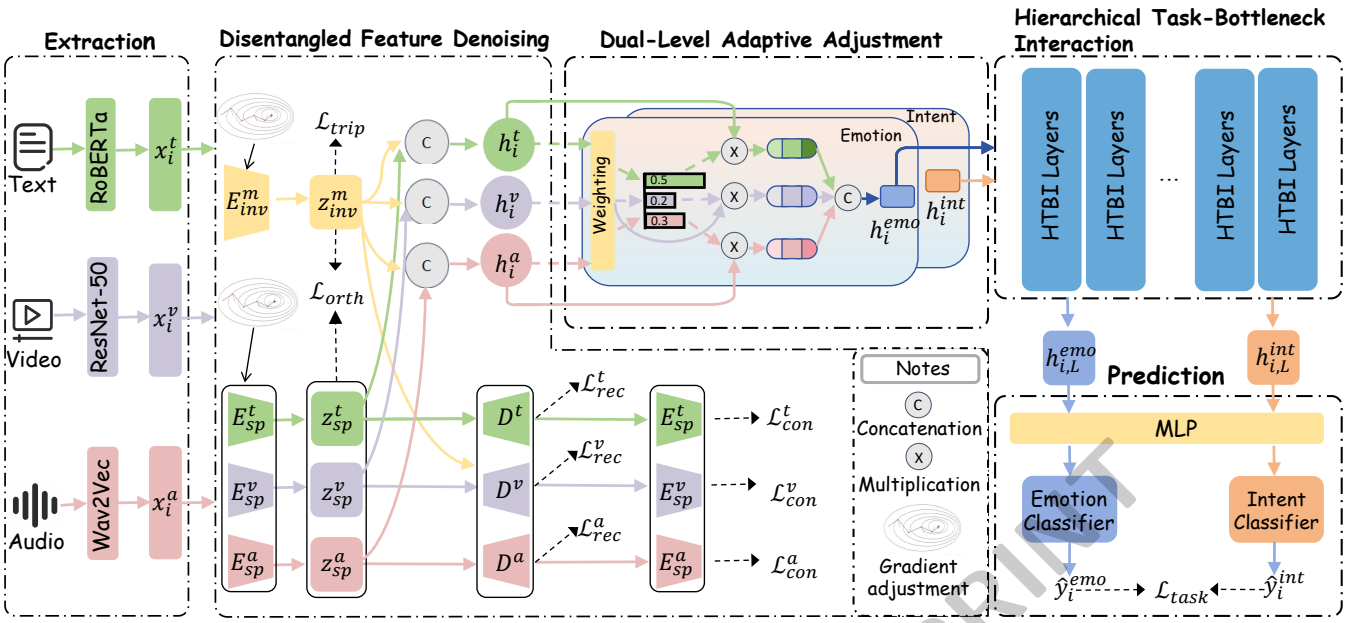

Figure 2: Overall architecture of the DELTA framework. It consists of Extracion, Disentangled Feature Denoising, Dual-Level Adaptive Adjustment, Hierarchical Task-Bottleneck Interaction, and Prediction modules. Features are initially purified to eliminate noise, dynamically adjusted for modality weights and sample difficulty, iteratively refined through the hierarchical bottleneck mechanism, and finally used to produce precise joint predictions.

from neighboring utterances to predict both labels. Building on this, [Liu *et al.*, 2024] introduced a novel framework, which explicitly models cross-modal interactions to capture the correlations between emotion and intent. More recently, [Yang *et al.*, 2025] presented the BEAR framework to handle modality uncertainty while exploiting task dependencies. However, despite these advancements, existing methods often treat multimodal inputs as clean signals. They may not thoroughly eliminate inherent noise within modalities and tend to overlook the varying contributions of different modalities to the final prediction. Furthermore, most approaches rely on shallow or direct interaction mechanisms. They often lack the sufficient capability to capture the deep, iterative mutual reinforcement between emotion and intent, which limits their ability to fully leverage the implicit and subtle correlations shared between these two tasks.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{D} = \{u_1, u_2, \ldots, u_n\}$ denote a multimodal dialogue consisting of n consecutive utterances, where speakers interact sequentially. For the i-th utterance u_i $(1 \leq i \leq n)$, the input involves three heterogeneous modalities: text (t), acoustic (a), and visual (v). We denote the extracted feature representations as $x_i^t \in \mathbb{R}^{d_t}$, $x_i^a \in \mathbb{R}^{d_a}$, and $x_i^v \in \mathbb{R}^{d_v}$, respectively, where d_m represents the feature dimension of modality $m \in \{t, a, v\}$. Each utterance u_i is associated with an emotion label y_i^{emo} and an intent label y_i^{int}, belonging to the predefined category sets $\mathcal{Y}_{emo}$ and $\mathcal{Y}_{int}$, respectively. Our objective is to jointly predict the associated emotion label y_i^{emo} and intent label y_i^{int} for every given utterance u_i.

3.2 Overview

Figure 2 illustrates the overall architecture of our DELTA framework. DELTA first employs a Disentangled Feature Denoising module to project heterogeneous inputs into orthogonal subspaces, effectively eliminating redundancy. Subsequently, a Dual-Level Adaptive Adjustment mechanism dynamically calibrates modality weights and modulates instance gradients. Finally, the Hierarchical Task-Bottleneck Interaction module serves as a progressive communication hub, facilitating deep, noise-filtered mutual enhancement between emotion and intent for precise joint prediction.

3.3 Disentangled Feature Denoising

Multimodal data often contains prevalent redundancy and conflicting information across modalities, which potentially hinders the accurate understanding of core emotions and intents. To address this challenge, following [Li *et al.*, 2023c], we propose a disentangled feature denoising module based on orthogonal decomposition. This module is designed to disentangle raw modal representations into modality-invariant subspaces and modality-specific subspaces, utilizing explicit constraints to reduce feature redundancy and filter out irrelevant noise.

For each input modal feature x_i^m $(m \in \{t, a, v\})$, we employ a shared E_{inv}^m encoder and three modality-specific encoders E_{sp}^m to disentangle it into modality-invariant features z_{inv}^m and modality-specific features z_{sp}^m:

$$z_{inv}^m = E_{inv}^m(x_i^m), \quad z_{sp}^m = E_{sp}^m(x_i^m), \tag{1}$$

where z_{inv}^m aims to capture the consistent emotion and intent semantics shared across modalities, while z_{sp}^m captures the unique information and variations specific to that modality.

To mitigate the redundancy between the two feature subspaces, we impose an orthogonality constraint by minimizing the cosine similarity between them, thereby forcing the model to learn complementary feature representations:

$$\mathcal{L}_{orth} = \sum_{m\in\{t,a,v\}} \cos(z^m_{inv}, z^m_{sp}). \tag{2}$$

To ensure that the disentangled features retain all essential information from the original input, we concatenate them and reconstruct the original feature via a decoder D^m. Simultaneously, to guarantee the consistency of the private features, the reconstructed feature is fed back into the private encoder to obtain $\tilde{z}^m_{sp}$:

$$\hat{x}^m_i = D^m(\text{Concat}(z^m_{inv}, z^m_{sp})), \quad \tilde{z}^m_{sp} = E^m_{sp}(\hat{x}^m_i). \tag{3}$$

The corresponding reconstruction loss $\mathcal{L}_{rec}$ and consistency loss $\mathcal{L}_{con}$ are defined as:

$$\mathcal{L}_{rec} = \|x^m_i - \hat{x}^m_i\|^2_2, \quad \mathcal{L}_{con} = \|z^m_{sp} - \tilde{z}^m_{sp}\|^2_2. \tag{4}$$

Although the above process achieves feature disentanglement, to further enhance the discriminability of the shared semantic feature z^m_{inv}, making samples with the same emotion label aggregate while separating those with different labels, we introduce a triplet loss:

$$\mathcal{L}_{trip} = \frac{1}{|R|} \sum_{(a,p,n)\in R} \max(0, \cos(z^a_{inv}, z^n_{inv}) - \cos(z^a_{inv}, z^p_{inv}) + \mu), \tag{5}$$

where R denotes the set of triplets, with (a, p, n) representing the anchor, positive, and negative samples respectively, and μ is the margin parameter.

The total loss $\mathcal{L}_{dec}$ for this module is calculated as the weighted sum of the aforementioned terms:

$$\mathcal{L}_{dec} = \alpha_o\mathcal{L}_{orth} + \alpha_r\mathcal{L}_{rec} + \alpha_c\mathcal{L}_{con} + \alpha_t\mathcal{L}_{trip}, \tag{6}$$

where α represents the balancing coefficients.

Finally, we fuse the purified modality-invariant feature with the modality-specific feature to obtain the final denoised feature h^m_i, serving as input for subsequent modules:

$$h^m_i = z^m_{inv} + z^m_{sp}. \tag{7}$$

Through such an orthogonal decomposition mechanism, we can effectively eliminate information redundancy and filter out irrelevant noise, while retaining key cues contributing to the understanding of emotion and intent. This process provides purer and more representative features for subsequent fusion and interaction.

3.4 Dual-Level Adaptive Adjustment

In multimodal dialogue understanding, the contribution of different modalities to emotion recognition and intent classification often varies dynamically with the context. Traditional methods typically treat all modalities equally, overlooking differences in their expressive capabilities, which may lead to a dominant modality overshadowing critical subtle cues from others. Furthermore, significant differences exist in learning difficulty across samples; the dominance of simple samples may limit the model's ability to generalize to hard samples. To address these issues, we propose an Dual-Level Adaptive Adjustment Mechanism that dynamically adjusts the learning process from two parts: Modality-level Dynamic Cooperation and Instance-level Gradient Adjustment.

Inspired by [He *et al.*, 2025], we first introduce Modality-level Dynamic Weighting (MDW) to achieve a refined, context-aware fusion of emotion and intent features. Unlike static concatenation, this strategy generates task-specific importance weights based on the joint multimodal context. Specifically, for the decoupled features h^m_i ($m \in \{t, a, v\}$), we first concatenate them to capture global interaction information. Subsequently, we utilize two independent task-specific Multi-Layer Perceptrons (MLP_{emo} and MLP_{int}) to project this context into importance scores $s^m_{i,\tau}$. These scores are then normalized via a Softmax function to obtain the final weights $\omega^m_{i,\tau}$:

$$s_{i,\tau} = \text{MLP}_\tau(\text{Concat}(h^t_i, h^a_i, h^v_i)), \quad \tau \in \{emo, int\} \tag{8}$$

$$\omega^m_{i,\tau} = \frac{\exp(s^m_{i,\tau})}{\sum_{k\in\{t,a,v\}} \exp(s^k_{i,\tau})}, \tag{9}$$

where $s^m_{i,\tau}$ denotes the logical score of modality m for task τ. The normalization ensures that $\sum_m \omega^m_{i,\tau} = 1$. Based on these weights, we perform weighted integration to obtain the final emotion feature h^{emo}_i and intent feature h^{int}_i:

$$h^{emo}_i = \sum_{m\in\{t,a,v\}} \omega^m_{i,emo} \cdot h^m_i, \quad h^{int}_i = \sum_{m\in\{t,a,v\}} \omega^m_{i,int} \cdot h^m_i. \tag{10}$$

Complementing forward fusion, we employ Instance-level Gradient Adjustment (IGA) to balance simple and challenging samples. This strategy dynamically modulates encoder parameter updates by monitoring prediction confidence. We first calculate the prediction deviation $\mathcal{D}^m_i$ for modality m to quantify the discrepancy between predictions and ground truth:

$$p^m_i = \text{Softmax}(\text{Proj}^m(h^m_i))_{I_i}, \quad \mathcal{D}^m_i = |1 - p^m_i|, \tag{11}$$

where I_i is the index of the true label for sample i, p^m_i represents the predicted probability for the target class, and Proj^m denotes the modality-specific projection layer. Next, we derive the adjustment weight W^m_i based on the relative deviation disparity across the batch:

$$W^m_i = \frac{1}{2} \sum_{k\neq m} \frac{\sum_{j\in\mathcal{B}, j\neq i}(\mathcal{D}^m_j - \mathcal{D}^k_j)}{\sum_{j\in\mathcal{B}}(\mathcal{D}^m_j - \mathcal{D}^k_j)}, \tag{12}$$

where k iterates over the other modalities distinct from m, $\mathcal{B}$ represents the current training batch, and j is the sample index. To apply this, we expand W^m_i into a weight matrix $\hat{W}^m_i \in \mathbb{R}^{d_m \times d_m}$ and modulate the gradient of the modality encoder parameters θ_m:

$$\theta^{(t+1)}_m = \theta^{(t)}_m - \eta \cdot (\hat{W}^m_i \odot \nabla_{\theta_m}\mathcal{L}_{total}). \tag{13}$$

In this way, we can dynamically adjust the optimization focus at the instance level and the feature contribution at the modality level. This dual-level adjustment enables the model to fully leverage the unique advantages of heterogeneous data while preventing any single modality or simple sample pattern from dominating the learning process. This can enhance the joint understanding ability of emotion and intent.

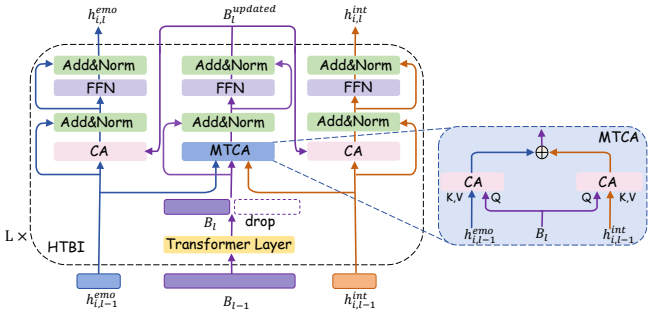

Figure 3: The structure of the HTBI module, which realizes mutual optimization through progressive bottleneck compression. In the diagram: B_{l-1} is compressed via drop to obtain the compact bottleneck B_l. MTCA aggregates task cues via parallel summation $\oplus$ to produce the refined hub $B_l^{updated}$. Finally, $h_{i,l}^{emo}$ and $h_{i,l}^{int}$ denote the updated emotion and intent features in the l-th layer, respectively, which retrieve enhanced context from the shared bottleneck.

3.5 Hierarchical Task-Bottleneck Interaction

Existing methods predominantly rely on single-round direct interactions, failing to account for the profound and iterative interplay between tasks. Consequently, they struggle to capture the intricate semantic dependencies between emotion and intent. Furthermore, traditional direct interaction methods often allow noise to propagate between tasks. To address these limitations, inspired by [Zheng *et al.*, 2025], we devise a Hierarchical Task-Bottleneck Interaction (HTBI) module. Specifically, we introduce a set of learnable "bottleneck tokens" acting as a communication hub to facilitate bidirectional information flow. Unlike simple direct interactions, the number of bottleneck tokens is progressively reduced at each layer. This design compresses and preserves critical interaction information and filters out irrelevant details layer-by-layer.

Specifically, as shown in Figure 3, the module consists of L iterative layers. In the l-th layer ($l \in \{1, \ldots, L\}$), we first determine the current bottleneck representation B_l. To achieve progressive information compression, the number of bottleneck tokens K_l is halved relative to the previous layer (i.e., $K_l = \lfloor K_{l-1}/2 \rfloor$). For the first layer, B_0 is initialized with learnable parameters. For subsequent layers, we employ a Transformer layer to contextualize the previous bottleneck features B_{l-1} and retain the top-K_l tokens:

$$\tilde{B}_{l-1} = \text{Transformer}(B_{l-1}), \tag{14}$$

$$B_l = \text{Truncate}(\tilde{B}_{l-1}, K_l), \tag{15}$$

where $\text{Transformer}(\cdot)$ denotes the standard encoder layer, and $\text{Truncate}(\cdot, K_l)$ represents the operation of slicing the first K_l tokens. This step acts as a semantic filter, producing a condensed summary of the task interaction history.

After obtaining the compressed bottleneck B_l, the interaction proceeds in two phases: aggregation and dissemination. We first formulate the standard Cross-Attention (CA) mechanism as: $\text{CA}(Q, K, V) = \text{Softmax}(\frac{QK^\top}{\sqrt{d_k}})V$. In the aggregation phase, the bottleneck tokens B_l act as the *Query*, seeking complementary information from the task features. As illustrated in the MTCA block of Figure 3, we employ a parallel interaction strategy. B_l independently queries the emotion features $h_{i,l-1}^{emo}$ and intent features $h_{i,l-1}^{int}$ from the previous layer. The retrieved contexts are then summed to integrate cues from both tasks:

$$\begin{aligned}\text{MTCA}(B_l, h_{i,l-1}^{emo}, h_{i,l-1}^{int}) = \text{CA}(B_l, h_{i,l-1}^{emo}, h_{i,l-1}^{emo}) \\ + \text{CA}(B_l, h_{i,l-1}^{int}, h_{i,l-1}^{int}).\end{aligned} \tag{16}$$

This aggregated information is fused into the bottleneck via a residual connection and normalization:

$$\hat{B}_l = \text{LayerNorm}(B_l + \text{MTCA}(B_l, h_{i,l-1}^{emo}, h_{i,l-1}^{int})) \tag{17}$$

$$B_l^{up} = \text{LayerNorm}(\hat{B}_l + \text{FFN}(\hat{B}_l)). \tag{18}$$

Through this parallel aggregation, the bottleneck hub effectively absorbs distinct yet complementary semantics from both emotion and intent domains without forcing them into a shared vector space prematurely.

Finally, in the dissemination phase, we utilize the refined bottleneck B_l^{up} to update the original task features. Here, the task-specific features (emotion or intent) act as the *Query*, while the updated bottleneck B_l^{upd} serves as the shared *Key* and *Value*. Each task retrieves enhanced interaction information from the common hub:

$$\tilde{h}_{i,l}^{\tau} = \text{LayerNorm}(h_{i,l-1}^{\tau} + \text{CA}(h_{i,l-1}^{\tau}, B_l^{up}, B_l^{up})), \tag{19}$$

$$h_{i,l}^{\tau} = \text{LayerNorm}(\tilde{h}_{i,l}^{\tau} + \text{FFN}(\tilde{h}_{i,l}^{\tau})), \quad \tau \in \{emo, int\} \tag{20}$$

Through L rounds of such hierarchical operations, the emotion and intent features continuously interact and refine each other. The final features not only retain their specific semantics but also integrate the filtered key information from the counterpart task, simulating the dynamic mutual influence observed in real conversations.

3.6 Prediction and Optimization

Following L rounds of hierarchical bottleneck interaction, the final obtained features $h_{i,L}^{emo}$ and $h_{i,L}^{int}$ not only retain their specific semantics but also fully integrate interaction information through the bottleneck hub. Finally, we feed these enhanced features into their respective MLPs to generate the final probability distributions for emotion and intent:

$$\hat{y}_i^{emo} = \text{Softmax}(\text{MLP}_e(h_{i,L}^{emo})), \tag{21}$$

$$\hat{y}_i^{int} = \text{Softmax}(\text{MLP}_i(h_{i,L}^{int})). \tag{22}$$

By employ the Cross-Entropy loss between the predicted results $\hat{y}$ and the ground truth labels y, the total task loss $\mathcal{L}_{task}$ is defined as the sum of the emotion prediction loss and the intent prediction loss:

$$\mathcal{L}_{task} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i^{emo}\log(\hat{y}_i^{emo}) + y_i^{int}\log(\hat{y}_i^{int})\right), \tag{23}$$

where N denotes the number of training samples. The final total loss function $\mathcal{L}_{total}$ is formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \mathcal{L}_{dec}. \tag{24}$$

Method	MCEIU_m				MCEIU_e				MINE			
	EMO		INT		EMO		INT		EMO		INT	
	ACC↑	WAF↑	ACC↑	WAF↑	ACC↑	WAF↑	ACC↑	WAF↑	ACC↑	WAF↑	M-F1↑	S-F1↑
bc-LSTM	48.91	44.74	49.55	54.76	35.28	40.11	38.94	40.05	59.12	57.88	47.33	49.65
MISA	50.48	48.51	50.09	56.26	36.74	41.46	40.22	38.11	63.28	58.76	52.27	52.54
MMIN	44.29	46.78	50.75	56.95	34.12	40.94	42.30	43.98	57.71	57.62	48.14	49.02
COGMEN	49.12	49.01	51.90	55.80	36.11	41.70	40.67	42.23	58.47	57.81	46.55	47.28
MFSA	45.67	47.23	49.81	51.44	34.55	38.77	41.05	39.88	64.05	57.19	**53.75**	53.01
Emolnt-Trans	45.76	50.47	52.05	58.45	34.78	40.40	42.95	44.46	59.33	56.74	47.12	48.69
JOYFUL	48.45	46.33	52.74	53.91	35.28	38.94	41.27	40.88	58.26	57.18	50.05	51.23
EI^2	**59.39**	55.08	61.97	61.63	47.84	42.09	51.12	45.53	62.74	59.63	52.33	53.56
BEAR	57.83	55.71	61.24	60.45	45.23	41.26	52.32	47.75	64.84	58.12	53.43	53.51
DELTA (Ours)	59.10	**58.61**	**65.70**	**63.50**	**52.71**	**47.58**	**57.02**	**51.12**	**65.82**	**61.97**	53.55	**54.36**

Table 1: Performance Comparison (%) with Baselines on MCEIU_m, MCEIU_e and MINE.

4 Experiments

4.1 Experimental Setup

Implementation Details. For feature extraction, we obtain 768-dimensional representations for all modalities. Specifically, audio clips are resampled to 16 kHz via Librosa and encoded by a pre-trained wav2vec 2.0 model [Baevski *et al.*, 2020]. Video frames are sampled at 1 Hz using FFmpeg and processed via the Video Swin Transformer [Liu *et al.*, 2022]. Textual features are derived from BERT-base [Devlin *et al.*, 2019]. During the training process, we select the Adam optimizer and a batch size of 32. We train the model for up to 100 epochs, implementing an early stopping strategy with a patience of 8. All experiments are conducted on NVIDIA A800 GPUs. Finally, we report the average results over five independent runs with different random seeds.

Datasets. We evaluate our proposed method on three datasets: MCEIU_m (Mandarin) [Liu *et al.*, 2024], MCEIU_e (English) [Liu *et al.*, 2024], and MINE [Yang *et al.*, 2025].

Evaluation Metrics. We evaluate performance using Accuracy (ACC) and Weighted Average F1-score (WAF). For the multi-label intent task in the MINE dataset, we additionally report Micro F1 (M-F1) and Samples F1 (S-F1).

4.2 Performance Comparisons

We compare DELTA against representative baselines, including bc-LSTM [Poria *et al.*, 2017], MISA [Hazarika *et al.*, 2020], MFSA [Yang *et al.*, 2022b], MMIN [Zhao *et al.*, 2021], COGMEN [Joshi *et al.*, 2022], JOYFUL [Li *et al.*, 2023b], EmoInt-Trans [Singh *et al.*, 2022], EI^2 [Liu *et al.*, 2024], and BEAR [Yang *et al.*, 2025]. From Table 1, we can have the following observations.

First, DELTA consistently outperforms baselines across all datasets. Specifically, on MCEIU_e, it achieves 52.71% and 57.02% accuracy for emotion and intent recognition. This is largely achieved by our DFD module, which employs orthogonal decomposition to rigorously decouple noise from essential semantics, overcoming the redundancy limitations of prior works.

Second, our model significantly outperforms methods that employ static fusion or treat modalities equally (e.g., MISA, MFSA). For instance, compared to MISA on MCEIU_m, DELTA improves the intent accuracy by over 15%. The reason might be that these baselines ignore the dynamic variation in modality contribution and sample difficulty, allowing dominant modalities or simple samples to bias the learning process. Conversely, DELTA employs a Dual-Level Adaptive Adjustment mechanism. This enables the model to dynamically re-weight modalities based on context and calibrate gradients for hard samples, ensuring robust understanding.

Third, while the unified frameworks EI^2 and BEAR also belong to the MC-EIU task, they still fall short compared to our method. For example, on MINE, DELTA improves the Intent S-F1 by 0.85% over BEAR. The reason might be that these methods primarily rely on shallow or direct interaction mechanisms, which fail to capture the deep iterative dependencies and risk propagating noise between tasks. Unlike these works, our HTBI module can act as a progressive communication hub, facilitating multi-round, noise-filtered bidirectional information flow, simulating the deep interplay between emotion and intent to achieve mutual enhancement.

4.3 Ablation Study

To thoroughly evaluate the effectiveness of the DELTA framework, we conduct ablation experiments on the MCEIU_m dataset, covering task integration, modality importance, and component contributions. The results are presented in Table 2.

First, regarding task integration and modalities, the proposed joint learning framework significantly outperforms single-task baselines with an improvement of 3.68% in Emotion Accuracy, confirming the necessity of leveraging task interdependence. Among modalities, text contributes the primary semantics, but removing audio or visual cues causes a notable performance drop, validating that non-verbal signals are indispensable for robust understanding. Second, we analyze the contribution of each module. Removing the Disentangled Feature Denoising (DFD) module incurs an accuracy loss of approximately 1.5%, verifying its role in filtering redundancy. Similarly, the exclusion of the Modality Dynamic Weighting (MDW) and Instance-Guided Adjustment (IGA) mechanisms leads to distinct performance declines. For instance, the absence of MDW reduces the Emotion Accuracy

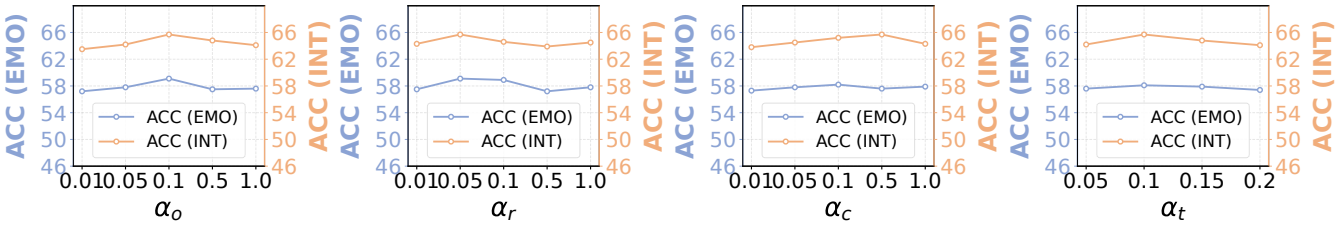

Figure 4: Sensitivity analysis of the key hyper-parameters (α_o, α_r, α_c, α_t) on the MCEIU_m dataset.

Method	EMO		INT	
	ACC↑	WAF↑	ACC↑	WAF↑
Single Task				
Emotion Task	55.42	53.15	–	–
Intention Task	–	–	62.18	59.85
Importance of Modalities				
A	26.45	23.89	30.12	28.56
V	24.33	22.17	27.58	26.94
L	57.34	56.55	63.89	61.72
A + V	28.91	26.43	33.05	30.22
L + A	58.05	57.29	64.66	62.33
L + V	57.92	57.48	64.31	62.15
Importance of Modules				
w/o DFD	57.65	56.88	64.12	62.05
w/o MDW	57.89	57.12	63.98	61.85
w/o IGA	57.45	56.95	64.45	62.38
w/o HTBI	57.20	56.50	63.55	61.40
DELTA (Ours)	**59.10**	**58.61**	**65.70**	**63.50**

Table 2: The results of ablation studies (%) on MCEIU_m.

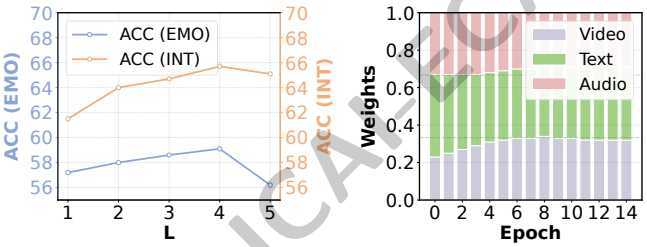

Figure 5: **Left:** The influence of the number of Hierarchical Task-Bottleneck Interaction (HTBI) layers L on emotion and intent recognition accuracy. **Right:** Visualization of the learned weights for different modalities (Video, Text, Audio) across training epochs.

to 57.89%, highlighting the importance of dynamically balancing modality weights. Finally, removing the Hierarchical Task-Bottleneck Interaction (HTBI) module causes the most significant impact among components, proving that our progressive bottleneck strategy is critical for capturing deep task interactions while preventing noise propagation.

4.4 Parameter and Weight Analysis

Sensitivity of Hyperparameters. We analyze the impact of the interaction depth L and the DFD balancing coefficients on model performance. As illustrated in the left panel of Figure 5 , accuracy improves steadily and peaks at $L = 4$, confirming that sufficient bottleneck iterations are essential for deep semantic refinement. However, a decline at $L = 5$ suggests that excessive depth may introduce redundancy and lead to overfitting. Furthermore, regarding the DFD module, we evaluate the loss weights $\alpha_o, \alpha_r, \alpha_c, \alpha_t$ in Figure 4. Empirical results indicate the model is robust to their variations. Balanced coefficients are found to be crucial: they ensure the orthogonality constraint ($\mathcal{L}_{orth}$) effectively filters noise without overwhelming the reconstruction ($\mathcal{L}_{rec}$) and primary task optimization, thereby guaranteeing feature purity and training stability.

Dynamic Modality Weights. The evolution of weights assigned by the MDW mechanism across training epochs is visualized in the right panel of Figure 5. On one hand, the weights exhibit a stable distribution after the initial phase, validating that the model converges to a robust fusion strategy. On the other hand, the weight of the text modality is consistently higher than that of audio and video. This showcases the intrinsic dominance of textual semantics in conversation, while highlighting the necessity of MDW. It ensures non-verbal modalities are not ignored but assigned appropriate attention to provide complementary cues, achieving superior fusion compared to static methods.

5 Conclusion

In this paper, we propose a novel framework for MC-EIU called Disentangled Hierarchical Interaction and Adaptive Adjustment. We introduce a Disentangled Feature Denoising module based on orthogonal decomposition to purify heterogeneous representations, and design a Dual-Level Adaptive Adjustment mechanism to dynamically optimize learning focus from both modality and instance perspectives. Furthermore, we employ a Hierarchical Task-Bottleneck Interaction strategy to facilitate progressive, noise-filtered mutual reinforcement between tasks. Experiments demonstrate that DELTA consistently outperforms state-of-the-art methods on standard benchmarks. Future work will explore incorporating Large Language Models to further enhance the semantic reasoning capabilities of the framework.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Nos. 62271411, U22A2098), the Technological Innovation Team of Shaanxi Province (No. 2025RS-CXTD-009), the International cooperation project of Shaanxi Province (No. 2025GH-YBXM-017).

References

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[Danieli *et al.*, 2015] Morena Danieli, Giuseppe Riccardi, and Firoj Alam. Emotion unfolding and affective scenes: A case study in spoken conversations. In *Proceedings of the International Workshop on Emotion Representations and Modelling for Companion Technologies*, pages 5–11, 2015.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[Fang *et al.*, 2025] Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14314–14324, 2025.

[Francese and Attanasio, 2023] Rita Francese and Pasquale Attanasio. Emotion detection for supporting depression screening. *Multimedia Tools and Applications*, 82(9):12771–12795, 2023.

[Gong *et al.*, 2022] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

[Hazarika *et al.*, 2020] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131, 2020.

[He *et al.*, 2025] Wen-Jue He, Xiaofeng Zhu, and Zheng Zhang. Cross-modal prompting for balanced incomplete multi-modal emotion recognition. *arXiv preprint arXiv:2512.11239*, 2025.

[Jiang *et al.*, 2025] Shenjie Jiang, Zhuoyu Wang, Xuecheng Wu, Hongru Ji, Mingxin Li, Xianghua Li, and Chao Gao. Ddse: A decoupled dual-stream enhanced framework for multimodal sentiment analysis with text-centric ssm. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5893–5902, 2025.

[Joshi *et al.*, 2022] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. Cogmen: Contextualized gnn based multimodal emotion recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, 2022.

[Li *et al.*, 2023a] Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5923–5934, 2023.

[Li *et al.*, 2023b] Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. Joyful: Joint modality fusion and graph contrastive learning for multimodal emotion recognition. *arXiv preprint arXiv:2311.11009*, 2023.

[Li *et al.*, 2023c] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640, 2023.

[Li *et al.*, 2024] Junlin Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. Be helpful but don't talk too much—enhancing helpfulness in conversations through relevance in multi-turn emotional support. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1976–1988, 2024.

[Liu *et al.*, 2022] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[Liu *et al.*, 2024] Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W Schuller, and Haizhou Li. Emotion and intent joint understanding in multimodal conversation: A benchmarking dataset. *arXiv preprint arXiv:2407.02751*, 2024.

[Liu *et al.*, 2026] Zhaowei Liu, Sheng Liu, Weiqing Yan, Peng Song, Yongchao Song, and Rufei Gao. Atgfb-mff: Adaptive text-guided fiber bundle feature fusion with llms for multimodal sentiment analysis and emotion recognition in conversations. In *Proceedings of the ACM Web Conference 2026*, pages 7442–7453, 2026.

[Poria *et al.*, 2017] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, 2017.

[Singh *et al.*, 2022] Gopendra Vikram Singh, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. Emointrans: A multimodal transformer for identifying emotions and intents in social conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:290–300, 2022.

[Tsai *et al.*, 2018] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

[Wu *et al.*, 2025] Ruo-Yan Wu, Xin-Heng Li, Yi-Chen Li, Zhi-Hong Ren, Bing-Xiang Yang, Zhen-Tao Liu, Bao-Liang Zhong, and Chen-Ling Liu. The effect of social robot interventions on anxiety in children in clinical settings: a systematic review and meta-analysis. *Journal of affective disorders*, 2025.

[Xia *et al.*, 2025] Wuyou Xia, Guoli Jia, Sicheng Zhao, and Jufeng Yang. Seek common ground while reserving differences: Semi-supervised image-text sentiment recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29601–29611, 2025.

[Yang *et al.*, 2022a] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1642–1651, 2022.

[Yang *et al.*, 2022b] Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. Learning modality-specific and-agnostic representations for asynchronous multimodal language sequences. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1708–1717, 2022.

[Yang *et al.*, 2024] Qu Yang, Mang Ye, and Dacheng Tao. Synergy of sight and semantics: Visual intention understanding with clip. In *European Conference on Computer Vision*, pages 144–160. Springer, 2024.

[Yang *et al.*, 2025] Qu Yang, Qinghongya Shi, Tongxin Wang, and Mang Ye. Uncertain multimodal intention and emotion understanding in the wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24700–24709, 2025.

[Zhao *et al.*, 2021] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. 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 2608–2618, 2021.

[Zheng *et al.*, 2025] Li Zheng, Tengyue Song, Yuzhe Ding, Xiaorui Wu, Fei Li, Dongdong Xie, Jinbo Li, Chong Teng, and Donghong Ji. Improving emotion and intent understanding in multimodal conversations with progressive interaction. *IEEE Transactions on Affective Computing*, pages 1–13, 2025.

[Zhu *et al.*, 2019] Yaochen Zhu, Zhenzhong Chen, and Feng Wu. Multimodal deep denoise framework for affective video content analysis. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 130–138, 2019.