Emotion-Wheel-Guided Audio-Referred Text Representation for Multimodal Emotion Recognition in Conversation

Eunseon Seong[1*], **Harim Lee**[1*], **Dahye Kim**[1], **Changhyun Kim**[2] and **Dong-Kyu Chae**[1†]
[1]Department of Artificial Intelligence, Hanyang University, Seoul, South Korea
[2]Metropolitan AT/DT Drive Team, SK Telecom, Seoul, South Korea
{emilyseong, hrimlee, dahye99, dongkyu}@hanyang.ac.kr
changhyk@sk.com

Abstract

Multimodal Emotion Recognition in Conversation aims to identify emotions within a dialogue with multimodal data, including audio, visual, and textual features. While existing methods have made significant improvements, there are two fundamental limitations to be addressed. From the modality fusion perspective, current approaches treat all modalities as functionally equivalent during fusion, overlooking their distinct communicative roles and information capacities, in which text conveys explicit semantic meaning while audio provides paralinguistic cues. From the emotion label perspective, many works ignore the continuous structure of emotion characterized by psychological theory and apply uniform penalties regardless of affective proximity. To address these limitations, we propose EMART, **EM**otion-Wheel-Guided **A**udio-**R**eferred **T**ext Representation for ERC, specifically focusing on audio and text modalities. First, we propose a modality-aware fusion strategy capturing linguistic features from text as the primary source and audio as a complementary component. Secondly, we propose an emotion-wheel-guided supervised contrastive loss to encode emotional proximity based on Russell's circumplex model. Experimental results on IEMOCAP and MELD demonstrate outstanding performance. The code is available at: https://github.com/DILAB-HYU/EMART.git.

1 Introduction

Emotion is intrinsic to human interaction, establishing emotion recognition as a key element of human-like artificial intelligence (Poria et al., 2019b; Liu et al., 2021; Tu et al., 2022; Li et al., 2019). Accordingly, Emotion Recognition in Conversation (ERC) has emerged as a central task in natural language processing, aiming to identify the emotion of each utterance within a dialogue. Traditional ERC approaches primarily relied on textual features (Majumder et al., 2019; Ghosal et al., 2019). More recently, multimodal strategies have been adopted to capture a broader spectrum of emotional expressions, by incorporating acoustic and visual signals alongside text (Zhang and Tan, 2025; Hu et al., 2021, 2022; Chen et al., 2023). While such multimodal approaches have demonstrated their superiority by incorporating information from multiple modalities, there remain two fundamental limitations to be addressed.

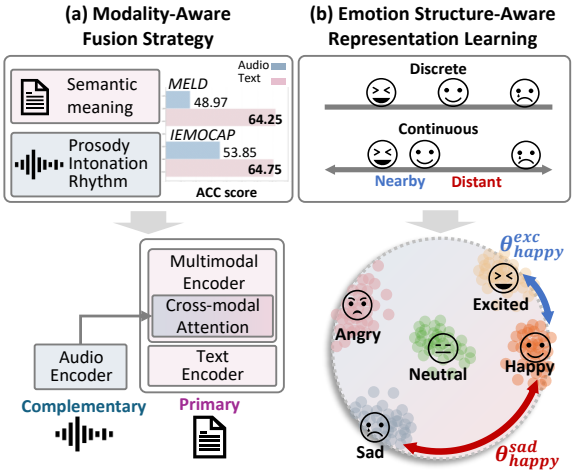


Figure 1: Main Idea of EMART. (a) Modality-aware fusion strategy, (b) Emotion-structure-aware representation learning.

First, existing multimodal ERC models typically treat all modalities as functionally equivalent during fusion, overlooking their distinct communicative roles and information capacities. These approaches integrate modalities using graph neural networks or bidirectional cross-attention mechanisms (Hu et al., 2021, 2022; Zhang and Tan, 2025), implicitly assuming that each modality contributes equally to emotion recognition. However, modalities play fundamentally distinct roles in emotional communication and differ in the degree of emotion-

*Equal contribution.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 40392–40403

related information they provide. From a linguistic perspective, text serves as the primary source of explicit semantic meaning, whereas audio conveys paralinguistic cues such as prosody, intonation, and rhythm (Schuller et al., 2013; Ephratt, 2011). Consistent with this distinction, the empirical results in Figure 1 show that text substantially outperforms audio in emotion recognition, reflecting their different information capacities. This observation suggests that treating modalities equally during fusion may be suboptimal.

Considering this discrepancy, we propose a **modality-aware fusion mechanism that accounts for the distinct roles and information capacities of different modalities**. Our work specifically focuses on two modalities, audio and text, capturing the linguistic feature from the text as the primary source of explicit meaning, and integrating paralinguistic cues derived from audio as a complementary component. In detail, we extract audio representations through an audio encoder and integrate them into the text encoder via cross-attention layers to obtain audio-referred text representations. To address the modality discrepancy, we introduce an alignment module before fusion, ensuring coherent multimodal integration.

Beyond effective modality fusion, a key limitation of existing ERC approaches lies in their failure to account for the inherent structure of emotion labels during training. Psychologically, emotion has been characterized along continuous dimensions of valence, arousal, and dominance, where semantically similar emotions (e.g., happy and excited) lie closer in affective space than distant ones (e.g., happy and sad). However, many studies overlook this continuous characteristic of emotions and solve the ERC task as a simple discrete classification. While recent work incorporates dimensional affect by mapping utterances to valence-arousal-dominance (VAD) space using fixed emotion prototypes (Yang et al., 2023) for emotion structure-aware representation learning, it applies conventional contrastive loss functions, treating all negative (i.e., different class) emotions uniformly, regardless of their affective distance. Consequently, the learned representation fails to reflect the proximity of emotion labels, penalizing misclassifications between similar emotions identically to misclassifications between opposite ones.

In this context, we introduce an **emotion-wheel-guided supervised contrastive objective that accounts for the emotional continuity grounded in psychological theory**. Specifically, we map emotions onto an emotion wheel represented by angular positions based on Russell's circumplex model (Russell, 1980), so that the distances reflect psychological proximity. Through an angle-based supervised contrastive loss, we encourage nearby emotions to be embedded closer while pushing distant ones apart proportional to their angular distances, aligning the representation with psychological structure.

To this end, we propose **EMART**, **EM**otion-Wheel-Guided **A**udio-**R**eferred **T**ext Representation for multimodal ERC, a framework that introduces an effective representation learning approach for this task.

Our contributions can be summarized as:

- We propose a modality-aware fusion architecture that treats text as the primary semantic modality and audio as complementary paralinguistic cues.
- We propose an emotion-wheel-guided supervised contrastive loss to encode emotional proximity, aligning learned representations with the psychological structure of emotions.

2 Related Works

2.1 Multimodal Emotion Recognition in Conversation

Recent multimodal ERC approaches treat modalities as functionally equivalent and focus on designing fusion mechanisms. MMGCN (Hu et al., 2021) constructs modality-specific graphs for contextual modeling, MM-DFN (Hu et al., 2022) employs dynamic fusion to reduce redundancy, M3Net (Chen et al., 2023) uses hypergraph propagation for multi-frequency relationships, and ECERC (Zhang and Tan, 2025) integrates emotional evidence with contextual causes through attention mechanisms. In contrast, CTAL (Li et al., 2021a) explicitly considers modality heterogeneity, treating audio as primary and incorporating text via cross-modal attention in audio-language pretraining. Recognizing the modality heterogeneity, we adopt a modality-aware fusion framework but treat text as the primary modality, integrating audio as complementary information.

2.2 Continuous Characteristic of Emotions

Most ERC studies formulate emotion recognition as a discrete classification problem (Hu et al., 2022;

Chen et al., 2023; Zhang and Tan, 2025), without accounting for the continuous relationships among emotion labels. To incorporate emotional continuity, SCCL (Yang et al., 2023) maps emotion clusters into the Valence-Arousal-Dominance (VAD) space, a dimensional affect model capturing pleasantness, activation, and control. They then perform cluster-level contrastive learning to incorporate measurable emotion prototypes. While this approach introduces emotional continuity into the representation space, it treats all negative pairs uniformly, pushing different emotion categories apart equally. In contrast, we propose an emotion-wheel-guided contrastive loss that explicitly aligns learned representations with emotional proximity

3 Method

Figure 2 illustrates the overall architecture of EMART for multimodal emotion recognition. This section is organized as follows: (3.1) Problem Statement, (3.2) Data Preprocessing, (3.3) Model Architecture, and (3.4) Training Objectives.

3.1 Problem Statement

Given a conversation consisting of N utterances, we define the set of utterances as $U = \{u_1, u_2, ..., u_N\}$ and the emotion labels $Y = \{y_1, y_2, ..., y_N\}$. Each utterance $u_i = \{a_i, t_i\}$ is represented by acoustic and textual modalities, a_i referring to the speech segment and t_i denoting the text transcript. The ERC task aims to predict the emotion label y_i of the corresponding utterance u_i.

3.2 Data Preprocessing

To capture conversational context, we prepend preceding utterances to the current one, providing contextual information for emotion recognition. In addition, we introduce two types of special tokens to emphasize the target utterance within long conversational histories: (1) **Speaker identifiers**: Since conversational dialogues often involve multiple speakers, we use <SELF> to indicate the current speaker and <OTHER> to denote other speakers. (2) **Utterance identifier**: We insert a [Current] token to explicitly mark the target utterance, enabling the model to focus on that utterance being classified. An example of the input construction process is illustrated in Table 1.

For the audio modality, we use the corresponding speech segment aligned with the current target text. The resulting audio input and text input are denoted X_a and X_t, respectively.

Utterance	Content
u_{t-1}	Hi. Where is everybody?
u_t	Oh, it's already closed, Chris gave me the keys to lock up, what is wrong?
X_t	<OTHER> Hi. Where is everybody? </s></s> [Current] <SELF> Oh, it's already closed, Chris gave me the keys to lock up, what is wrong?

Table 1: Example of input construction for the target utterance u_t. Preceding utterance u_{t-1} is prepended with <OTHER> and separated by </s></s>, while the target utterance u_t is marked with [Current] and <SELF>. Note that </s></s> stands for separator following the setting of RoBERTa (Liu et al., 2019).

3.3 Model Architecture

As the relative importance of modalities differs by task (Li et al., 2021a), prior works have demonstrated that unimodal encoders (e.g., RoBERTa) have strong potential to be extended into multimodal encoders through appropriate fusion mechanisms (Li et al., 2021b).

Motivated by this observation, we extend RoBERTa (Liu et al., 2019), a widely used encoder-only text model, into a multimodal encoder to integrate text and audio modalities. Specifically, we divide the $\text{RoBERTa}_{\text{base}}$ model into two parts: the first six layers serve as a unimodal text encoder, while the remaining six layers are used as a multimodal encoder enhanced with a cross-modal attention mechanism. The text encoder is initialized with the first six layers of the pretrained $\text{RoBERTa}_{[:6]}$ model, and the multimodal encoder is initialized with the last six layers of the pretrained $\text{RoBERTa}_{[6:12]}$ model. For audio, we use a pretrained model such as WavLM or Wav2vec to extract acoustic representations.

3.3.1 Unimodal Encoder

Given audio inputs X_a and text inputs X_t, we obtain their representations from the corresponding unimodal encoders:

$$\mathbf{H}_a = \textit{AudioEncoder}(X_a) \in \mathbb{R}^{B\times M\times d_a},$$
$$\mathbf{H}_t = \textit{TextEncoder}(X_t) \in \mathbb{R}^{B\times L\times d_t},$$

where L and M denote the text sequence length and the number of audio frames, respectively, and d_t, d_a are the hidden dimensions of each modality.

3.3.2 Multimodal Encoder

Following our design choice of a modality-aware fusion strategy, in which text serves as the primary semantic modality while audio provides complementary paralinguistic cues, we regard the textual

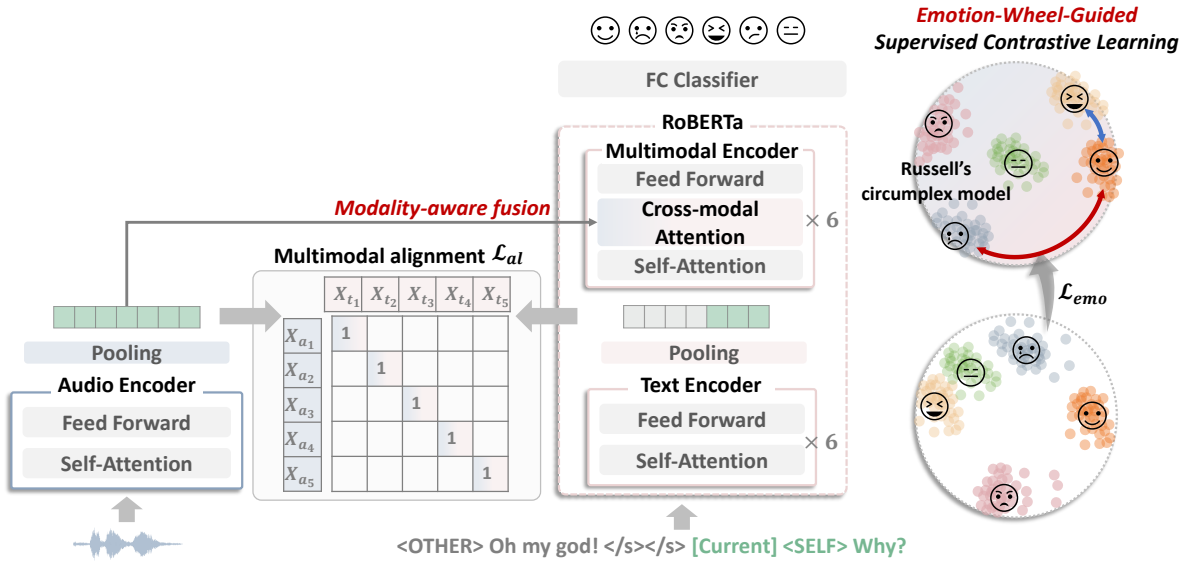


Figure 2: The overview of **EMART**, which is composed of an audio encoder, a text encoder, and a multimodal encoder, trained with an audio–text alignment loss and an emotion wheel–guided supervised contrastive loss.

representation as the main stream and incorporate audio in an auxiliary manner. Specifically, we introduce a cross-modal attention mechanism within the multimodal encoder. Each block in the multimodal encoder is computed as:

$$\tilde{\mathbf{H}}_t^\ell = \mathit{SelfAttn}(\mathbf{H}_t^{\ell-1}, \mathbf{H}_t^{\ell-1}, \mathbf{H}_t^{\ell-1}),$$
$$\hat{\mathbf{H}}_t^\ell = \mathit{CrossAttn}(\tilde{\mathbf{H}}_t^\ell, \mathbf{H}_a, \mathbf{H}_a),$$
$$\mathbf{H}_t^\ell = \hat{\mathbf{H}}_t^\ell + \mathit{FFN}(\hat{\mathbf{H}}_t^\ell),$$

where ℓ denotes the layer index in the multimodal encoder, and $\mathbf{H}_a$ represents the last hidden states of the audio encoder. Finally, the multimodal encoder yields the fused representation $\mathbf{H}$, where linguistic features are enriched by paralinguistic cues from the audio stream, reflecting the distinct communicative roles and information capacities of each modality in emotion recognition.

3.4 Training Objectives

In this section, we outline the training objective designed to optimize the proposed model. Beyond the standard Cross-Entropy Loss for classification, we incorporate two auxiliary objectives: (1) **Audio-Text Alignment Loss** for cross-modal alignment and (2) **Emotion-Wheel-Guided Supervised Contrastive Loss** for emotion-structure-aware representation learning.

3.4.1 Audio-Text Alignment Loss Before Fusion

To alleviate the discrepancy between the acoustic embeddings and textual embeddings derived from each encoding model, we employ an alignment module to ensure semantic alignment before fusion. For this purpose, we regard the audio and text embeddings from the same utterance as positive pairs and adopt the Barlow Twins loss (Zbontar et al., 2021) to maximize the correlation of positive pairs while reducing redundancy across dimensions.

Given unimodal representations $\mathbf{H}_a$ and $\mathbf{H}_t$ for audio and text, we mean pool $\mathbf{H}_a$ over the audio stream and $\mathbf{H}_t$ over the target utterance to obtain utterance-level representations $\overline{\mathbf{H}}_a \in \mathbb{R}^{B\times d_a}$ and $\overline{\mathbf{H}}_t \in \mathbb{R}^{B\times d_t}$. We then apply a single linear projection layer to each modality, $f_{\mathrm{a}}(\cdot)$ for acoustic embeddings and $f_{\mathrm{t}}(\cdot)$ for textual embeddings, projecting them into the same dimensional space with a size of d, such that

$$\mathbf{Z}_a = f_{\mathrm{a}}(\overline{\mathbf{H}}_a), \mathbf{Z}_t = f_{\mathrm{t}}(\overline{\mathbf{H}}_t),$$

where $\mathbf{Z}_a, \mathbf{Z}_t \in \mathbb{R}^{B\times d}$.

Next, we compute the cross-correlation matrix $C \in \mathbb{R}^{d\times d}$ between the acoustic and textual embedding as:

$$C_{jk} = \frac{\sum_{b=1}^{B} \mathbf{Z}_{a_b}^j \, \mathbf{Z}_{t_b}^k}{\sqrt{\sum_{b=1}^{B} (\mathbf{Z}_{a_b}^j)^2} \sqrt{\sum_{b=1}^{B} (\mathbf{Z}_{t_b}^k)^2}}, \quad (1)$$

where b denotes the batch index and j, k correspond to feature dimensions.

Algorithm 1: Training Procedure

Input: Audio-Text Paired Input $\{X_{t_i}, X_{a_i}\}_{i=1}^{N}$, Emotion labels $Y = \{y_i\}_{i=1}^{N}$, Audio encoder *AudioEnc*$(\cdot)$, Text encoder *TextEnc*$_{[:6]}(\cdot)$, Multimodal encoder *TextEnc*$_{[6:12]}(\cdot)$, Projectors $f_{\text{a}}(\cdot), f_{\text{t}}(\cdot), f_{\text{fuse}}(\cdot)$, Classification head $f_{\text{cls}}(\cdot)$

for *epoch* $= 1, \ldots, E$ **do**

  for *mini-batch* $\{X_{t_i}, X_{a_i}\}_{i=1}^{B}$ **do**

    $\mathbf{H}_a = \textit{AudioEnc}(X_a)$

    $\mathbf{H}_t = \textit{TextEnc}_{[:6]}(X_t)$

    $\overline{\mathbf{H}}_a, \overline{\mathbf{H}}_t = \text{MeanPool}(\mathbf{H}_a, \mathbf{H}_t)$

    $\mathbf{Z}_a = f_{\text{a}}(\overline{\mathbf{H}}_a)$

    $\mathbf{Z}_t = f_{\text{t}}(\overline{\mathbf{H}}_t)$

    {Audio-Text Alignment Loss}

    Compute $\mathcal{L}_{al}$ (Eq. 2)

    $\mathbf{H}_{\text{fuse}} = \textit{TextEnc}_{[6:12]}(\mathbf{H}_a, \mathbf{H}_t)$

    $\mathbf{Z}_{fuse} = f_{\text{fuse}}(\mathbf{H}_{fuse})$

    {Emotion-Wheel-Guided Supervised Contrastive Loss}

    Compute $\mathcal{L}_{emo}$ (Eq. 12)

    Compute final objective $\mathcal{L}$ (Eq. 13)

Finally, we define the audio-text alignment loss $\mathcal{L}_{al}$ as:

$$\mathcal{L}_{al} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2, \quad (2)$$

where the first term enforces cross-modal invariance and the second term reduces redundancy across other utterances. Here, a positive constant λ controls the importance of the two terms.

3.4.2 Emotion-Wheel-Guided Contrastive Alignment for Cross-Modal Fusion

Motivated by the continuous structure of emotions in psychological theory, we employ an emotion-wheel-guided supervised contrastive loss to encode the proximity among emotion labels.

Positive and Negative Pair Selection. We first define how we construct the positive and negative pairs for supervised contrastive learning for emotion recognition. While the selection of positive and negative pairs is generally straightforward for most emotions, *neutral* emotion needs special consideration. In the circumplex model, *neutral* is positioned near the origin, weakly related to all directional emotions. Consequently, we treat the *neutral* emotion as a 'weak positive' for any other emotion class. Here, $\mathcal{P}(i)$, $\mathcal{W}(i)$, and $\mathcal{N}(i)$ denote positive, weak positive, and negative sets for anchor i, respectively.

Emotion-Wheel-Guided Supervised Contrastive Loss. While conventional supervised contrastive learning (Khosla et al., 2020) is effective to separate all negative pairs (i.e., samples with different labels) uniformly, it fails to account for their underlying semantic similarity. Therefore, inspired by ACCon (Zhao et al., 2025), we modulate the strength of separation based on emotional proximity. This ensures that semantically similar emotions (e.g., *sad* and *frustrated*) are embedded closer together than dissimilar ones (e.g., *sad* and *excited*), preserving the continuous nature of the emotion space.

First, we map each emotion onto Russell's circumplex model as $\theta \in [0, 2\pi)$, where θ_i denotes the angular position of the emotion label for the sample u_i on the emotion wheel. For an anchor i and a negative sample m, the *true angular distance* on the emotion wheel is defined as:

$$\theta^*_{i,m} = \min\left(|\theta_i - \theta_m|, 2\pi - |\theta_i - \theta_m|\right), \quad (3)$$

Given the fused representation $\mathbf{H}$ and a single linear projection layer $f_{\text{fuse}}(\cdot)$, we obtain L2-normalized features $\mathbf{Z} = f_{\text{fuse}}(\mathbf{H})$. Let $\hat{\theta}_{i,m}$ denote the angle between $\mathbf{Z}_i$ and $\mathbf{Z}_m$, derived from their cosine similarity:

$$\cos(\hat{\theta}_{i,m}) = \mathbf{Z}_i^\top \mathbf{Z}_m. \quad (4)$$

In conventional supervised contrastive learning, negative pairs are uniformly pushed toward an opposite direction, i.e., $\hat{\theta}_{i,m} \to \pi$. To align the embedding with the *true angular distance* $\theta^*_{i,m}$, we introduce an *adjusted angle* ϕ as:

$$\tilde{\theta}_{i,m} = \hat{\theta}_{i,m} + \phi, \qquad \phi := \pi - \theta^*_{i,m}, \quad (5)$$

and optimize $\tilde{\theta}_{i,m} \to \pi$, ultimately enforcing $\hat{\theta}_{i,m} \to \theta^*_{i,m}$.

Based on Eq.5, we derive the **adjusted cosine similarity** (see Appendix A) as:

$$\cos(\tilde{\theta}_{i,m}) = (\mathbf{Z}_i^\top \mathbf{Z}_m)\cos(\phi) - |\sin(\phi)|\sqrt{1 - (\mathbf{Z}_i^\top \mathbf{Z}_m)^2 + \epsilon}, \quad (6)$$

where ϵ is a small constant for numerical stability. The adjusted cosine similarity effectively ensures

the alignment of negative pairs according to the semantic relationship in the emotion wheel.

We define three types of similarity scores corresponding to the relationships between emotion labels. For positive pairs sharing identical emotion labels, weak-positive pairs involving *neutral*, and negative pairs with the others, we compute $s^{+}_{(p,q)}, s^{\circ}_{(p,q)}$, and $s^{-}_{(p,q)}$ as:

$$s^{+}_{(p,q)} = \exp(\mathbf{Z}_p^{\top}\mathbf{Z}q/\tau) \quad (7)$$

$$s^{\circ}_{(p,q)} = \exp(\mathbf{Z}_p^{\top}\mathbf{Z}q/\tau') \quad (8)$$

$$s^{-}_{(p,q)} = \exp(\cos(\tilde{\theta}_{p,q})/\tau), \quad (9)$$

where τ and τ' are temperature parameters for positive/negative pairs and weak positives, respectively.

Finally, we derive the emotion-wheel-guided supervised contrastive loss for a sampled batch of size B, as shown in Eq.12.

$$\psi_i = \sum_{\substack{p\in\mathcal{P}(i)\\ w\in\mathcal{W}(i)}} \log \frac{s^{+}_{(i,p)} + \lambda_w s^{\circ}_{(i,w)}}{\left(\begin{array}{c}\sum_{q\in\mathcal{P}(i)} s^{+}_{(i,q)} + \lambda_w \sum_{r\in\mathcal{W}(i)} s^{\circ}_{(i,r)} \\ + \sum_{m\in\mathcal{N}(i)} s^{-}_{(i,m)}\end{array}\right)} \quad (10)$$

$$\mathcal{L}_{emo_i} = -\frac{\psi_i}{|\mathcal{P}(i)| + \lambda_w|\mathcal{W}(i)|} \quad (11)$$

$$\mathcal{L}_{emo} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{emo_i} \quad (12)$$

Here, λ_w represents the weak-positive scaling parameter and $|\cdot|$ indicates set cardinality.

3.5 Final Objective

The fused feature representation $\mathbf{H}$ is passed through a single linear classification head $f_{\text{cls}}(\cdot)$ to predict the emotion label as $\hat{y} = f_{\text{cls}}(\mathbf{H})$. For this classification task, we use a standard cross-entropy loss $\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i$.

Based on the two auxiliary objectives mentioned in the previous section, our final objective $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{cls} + \lambda_{al}\mathcal{L}_{al}, & \text{epoch} \le t \\ \mathcal{L}_{cls} + \lambda_{al}\mathcal{L}_{al} + \lambda_{emo}\mathcal{L}_{emo}, & \text{otherwise,} \end{cases} \quad (13)$$

with a hyperparameter λ_{al} controlling the importance of the audio-text alignment loss and λ_{emo} controlling the emotion-wheel-guided supervised contrastive loss. In the early stage, the model is trained with $\mathcal{L}_{cls}$ and $\mathcal{L}_{al}$ to attain alignment, and after t epochs, $\mathcal{L}_{emo}$ is incorporated.

	IEMOCAP			MELD		
	train	dev	test	train	dev	test
# Dialogues	108	12	31	1038	114	280
# Utterances	5163	647	1623	9989	1109	2610
# Classes	6 classes			7 classes		

Table 2: Number of dialogues and utterances of the IEMOCAP and MELD datasets.

IEMOCAP		MELD	
Neutral	-	Neutral	-
Sad	$9/8\pi$	Sad	$9/8\pi$
Frustrated	$7/8\pi$	Anger	$6/8\pi$
Angry	$6/8\pi$	Joy	$2/8\pi$
Happy	$1/8\pi$	Surprise	$4/8\pi$
Excited	$3/8\pi$	Fear	$5/8\pi$
		Disgust	$7/8\pi$

Table 3: Emotion-to-angle mapping on the circumplex model.

4 Experiments

4.1 Datasets

We evaluate EMART on two representative ERC datasets: IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019a). The overall statistics of both datasets are summarized in Table 2. **IEMOCAP** is a multimodal (audio, text, and visual) dyadic conversation dataset consisting of 7,433 utterances across 151 dialogues from five sessions. Each utterance is annotated with one of six emotions: happy, sad, neutral, angry, excited, or frustrated. We use the first three sessions for training, the fourth session for validation, and the fifth session for testing, following prior works. **MELD** is a multimodal (audio, text, and visual) multi-party conversation dataset, containing 1,432 dialogues and 13,708 utterances from 304 speakers. Each utterance is labeled with one of seven emotions: anger, disgust, sadness, joy, neutral, surprise, or fear. MELD provides official data splits, including 1,038 dialogues (9,989 utterances) for training, 114 dialogues (1,109 utterances) for validation, and the remaining dialogues for testing.

Table 3 shows the angular mapping of each dataset's emotion categories based on Russell's circumplex model (Russell, 1980; Feldman Barrett and Russell, 1998; Posner et al., 2005). As *neutral* is a baseline state rather than a specific emotional axis, we do not explicitly assign an angle.

Models	Emotion Categories of **IEMOCAP**						**Overall (weighted)**	
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc.	F1.
DialogueRNN	31.25	63.67	55.21	58.82	74.25	64.30	60.38	60.36
DialogueGCN	28.47	62.45	48.44	50.00	73.91	62.47	56.87	56.63
MMGCN	26.39	74.29	68.49	68.82	75.25	60.37	65.00	64.44
MM-DFN	24.31	75.10	63.54	71.18	73.58	69.29	65.13	64.65
M3Net	68.75	73.06	65.36	68.24	67.89	64.30	67.34	67.88
CTAL-base	35.42	58.78	67.97	55.29	36.45	59.32	54.53	54.44
CTAL-large	40.28	57.55	52.08	64.71	44.48	57.22	52.99	53.36
ECERC	57.64	78.78	66.67	74.12	69.57	62.99	68.15	68.36
EMART (WavLM)	60.42	79.59	83.07	61.17	66.22	64.57	**70.79**	**70.93**
EMART (Wav2vec)	52.78	80.41	69.01	62.94	68.23	68.50	**68.39**	**68.61**

	Emotion Categories of **MELD**							**Overall (weighted)**	
	Neutral	Surprise	Fear	Sad	Joy	Disgust	Angry	Acc.	F1.
DialogueRNN	77.15	42.35	0.00	25.48	51.49	0.00	56.23	59.08	57.55
DialogueGCN	77.55	21.35	0.00	7.69	44.78	0.00	50.72	53.83	51.00
MMGCN	79.38	61.21	0.00	11.54	58.21	0.00	32.17	58.93	56.03
MM-DFN	80.49	51.96	0.00	19.71	50.50	0.00	48.12	59.31	57.24
M3Net	84.24	60.85	8.00	28.37	65.42	23.53	42.32	65.79	64.12
CTAL-base	74.04	50.89	10.00	19.23	39.55	14.71	25.51	52.68	50.87
CTAL-large	77.15	54.80	6.00	23.08	37.81	0.00	22.03	53.72	50.65
ECERC	82.40	60.85	12.00	30.29	62.44	23.53	48.12	65.44	64.26
EMART (WavLM)	88.61	48.04	14.00	27.40	58.71	8.82	51.59	**66.36**	64.19
EMART (Wav2vec)	87.74	46.62	20.59	27.88	60.45	4.41	52.12	**66.70**	**64.71**

Table 4: Performance comparison on IEMOCAP and MELD. We report the best run across three random seeds, with mean and standard deviation reported in Appendix Table B.2. **F1.** denotes F1-score, and **Acc.** denotes Accuracy. All results are reproduced using the authors' original implementations. **EMART (WavLM)** denotes our model using WavLM as the audio encoder, while **EMART (Wav2vec)** denotes the variant using Wav2vec. For CTAL, experiments are conducted with two model sizes: CTAL-base and CTAL-large.

4.2 Baselines

We compare EMART against seven representative open-source ERC baselines, including: **Text-based models**: DialogueRNN and DialogueGCN, **Graph-based multimodal models**: MMGCN, MM-DFN, and M3Net; **Transformer-based multimodal models**: CTAL and ECERC. All baselines are reproduced using their official implementations. For multimodal baselines that originally utilize three modalities (audio, text, and visual), including MMGCN, MM-DFN, M3Net and ECERC, we re-implement them with the audio and text modalities for fair comparison.

4.3 Experimental Settings

Our model consists of a RoBERTa-base text encoder (125M parameters) and an audio encoder, which is either WavLM-base or Wav2vec-base (each with 94.4M parameters). The total number of parameters in our model is 229.6M. The hyperparameters of our model are selected via grid search. The model is trained for 20 epochs on IEMOCAP and 10 epochs on MELD, with learning rates of 2e-5 and 1e-5, and batch sizes of 64 and 32, respectively, using the Adam optimizer. All experiments are conducted on a single NVIDIA GeForce H100 GPU, with the average training times being 20 minutes for IEMOCAP and 10 minutes for MELD.

4.4 Main Results

Table 4 presents the performance comparison between our method and baseline models on IEMOCAP and MELD. Our approach demonstrates superior performance across both datasets in terms of Accuracy and F1-score.

Specifically, on IEMOCAP, EMART surpasses the previous state-of-the-art baseline (ECERC) by achieving 2.43% (WavLM) and 0.03% (Wav2vec) higher accuracy, as well as 2.78% (WavLM) and 0.46% (Wav2vec) higher F1-scores. On MELD,

our model yields 0.61% (WavLM) and 0.95% (Wav2vec) improvements in accuracy over ECERC. These results demonstrate the effectiveness of our model design.

Moreover, compared with CTAL, which also adopts a modality-aware fusion strategy but treats audio as the primary modality and text as a complementary cue, EMART achieves substantially higher performance, with gains of 16.01% (WavLM) and 13.61% (Wav2vec) on accuracy, and 15.93% (WavLM) and 13.61% (Wav2vec) in F1-scores on the IEMOCAP dataset. A similar performance trend is consistently observed on the MELD dataset. This highlights the advantage of our strategy, prioritizing text as the primary modality and integrating audio as complementary information.

Overall, these results confirm the effectiveness of EMART in extracting audio-referred text representation with an audio-text alignment module and capturing the emotional proximity via the emotion-wheel-guided contrastive learning module.

4.5 Analysis

In this section, we conduct analyses to examine the effectiveness of key components in our framework. WavLM is utilized for IEMOCAP and Wav2vec for MELD as the audio encoder, showing the best performance on their respective datasets.

Effectiveness of Fusion Strategy. In Table 5, we compare our multimodal fusion strategy with three representative fusion methods: concatenation (Concat), dot-product fusion (Dot Product), and bi-directional cross-modal attention (Attention), across three text encoders: BERT-base, ELECTRA-base, and RoBERTa-base.

With unimodal performance as a reference, RoBERTa consistently outperforms the unimodal audio baseline on both datasets, whereas BERT and ELECTRA underperform audio on IEMOCAP despite competitive results on MELD. This discrepancy reveals differences in information capacity across text encoders for capturing emotion-related features, which is reflected in the fusion results. When textual representations are weaker than audio (e.g., BERT/ELECTRA on IEMOCAP), bidirectional cross-attention performs best among fusion methods, as it allows the stronger audio stream to compensate for the weaker text representation. In contrast, when the text encoder provides more informative representations (e.g., RoBERTa on IEMOCAP and all encoders on MELD), our modality-aware fusion strategy consistently achieves the best performance among all fusion methods. These results demonstrate the effectiveness of our fusion strategy, which preserves informative text representations while selectively integrating complementary acoustic cues. Given recent advances in pre-trained language models, our text-centric approach becomes increasingly effective.

		IEMOCAP		MELD	
		Acc.	F1.	Acc.	F1.
Unimodal	Audio	53.85	52.46	48.97	38.05
	Text-BERT	44.49▼	44.45▼	58.24▲	53.42▲
	Text-ELECTRA	44.98▼	44.73▼	57.20▲	52.79▲
	Text-RoBERTa	64.70▲	64.54▲	64.75▲	64.27▲
Multimodal	*BERT*				
	Concat	58.16	57.79	56.17	53.67
	Dot product	60.26	59.87	57.05	54.44
	Attention	68.08	67.72	59.31	58.37
	Our strategy	64.70▼	64.67▼	61.99▲	59.64▲
	ELECTRA				
	Concat	57.30	56.90	52.76	48.27
	Dot product	58.53	58.61	56.44	52.04
	Attention	64.51	64.61	60.42	58.37
	Our strategy	64.02▼	63.88▼	61.42▲	59.75▲
	RoBERTa				
	Concat	64.51	64.39	62.61	60.75
	Dot product	66.36	66.21	63.87	62.21
	Attention	66.24	66.07	62.99	61.70
	Our strategy	70.79▲	70.93▲	66.70▲	64.71▲

Table 5: Effect of our fusion strategy. ▼ denotes that the text-based model performs worse than the corresponding audio-based model, whereas ▲ denotes that the text-based model performs better than the audio-based model.

Effectiveness of Emotion-Wheel-Guided Supervised Contrastive Loss. We analyze the effectiveness of our emotion-wheel-guided contrastive loss, designed to capture the psychological proximity among emotion labels based on their continuous affective structure. Figure 3 compares contrastive

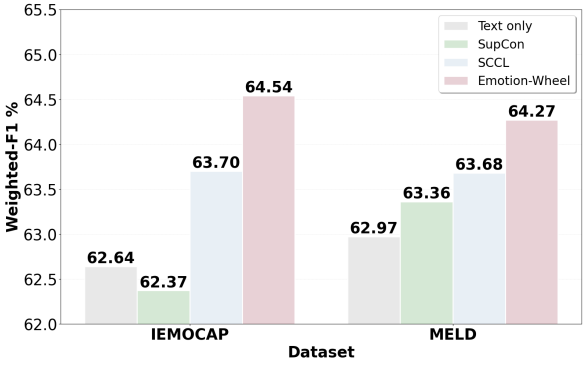


Figure 3: Effect of emotion-wheel mapping in supervised contrastive loss.

$\mathcal{L}_{al}$	$\mathcal{L}_{emo}$	IEMOCAP Acc.	IEMOCAP F1.	MELD Acc.	MELD F1.
✗	✗	64.70	64.58	63.10	62.31
✓	✗	68.02	67.06	65.25	63.30
✗	✓	68.21	68.21	65.75	63.72
✓	✓	**70.79**	**70.93**	**66.70**	**64.71**

Table 6: Ablation study on IEMOCAP and MELD.

loss variants (ours, SupCon, SCCL) on the text unimodal model, where 'Text-only' denotes the model without any contrastive loss. While SupCon can lead to performance degradation due to uniformly separating negative pairs, SCCL yields performance improvement by leveraging affective supervision. Our method further improves performance by explicitly modeling emotional proximity.

Ablation Study on Loss Terms. Table 6 presents an ablation study on $\mathcal{L}_{al}$ and $\mathcal{L}_{emo}$ to evaluate the contribution of each loss component. Compared to the model that applies neither loss, using either loss alone improves performance, while combining both yields the best results, demonstrating the effectiveness of each loss component. In particular, the emotion-wheel-guided supervised contrastive loss $\mathcal{L}_{emo}$ demonstrates a larger performance gain, underscoring the importance of emotion-structure-aware representation learning.

Effect of Fusion Start Layer. In Table 7, we examine the model performance on varying fusion start layers. Early stage fusion (1st and 3rd layers) disrupts the text encoder before stable linguistic representations are formed, resulting in the lowest performance. Conversely, late fusion (9th layer) leaves insufficient layers for meaningful cross-modal interaction, leading to performance degradation despite well-formed text representations. Our configuration (6th layer) achieves the best performance, with well-formed text representations and sufficient subsequent layers to effectively enable multimodal integration.

Layer	IEMOCAP Acc.	IEMOCAP F1.	MELD Acc.	MELD F1.
0	63.59	63.70	59.20	54.33
3	66.97	67.25	62.91	60.31
6	**70.79**	**70.93**	**66.70**	**64.71**
9	68.21	68.24	65.10	62.97

Table 7: Performance comparison on varying fusion start layer.

4.6 Hyperparameter Sensitivity

We report the hyperparameter sensitivity of the loss coefficients λ_w, λ_{al}, and λ_{emo} in Tables 8, 10, and 9, respectively. The results indicate that these hyperparameters are mutually entangled in learning feature representations.

λ_w	IEMOCAP Acc.	IEMOCAP F1.	MELD Acc.	MELD F1.
0.1	69.69	69.65	65.02	63.46
0.2	70.73	70.30	**66.70**	**64.71**
0.3	**70.79**	**70.93**	65.06	63.91
0.4	69.32	68.93	64.79	63.38
0.5	68.70	68.56	64.52	63.14

Table 8: Hyperparameter sensitivity on λ_w.

λ_{emo}	IEMOCAP Acc.	IEMOCAP F1.	MELD Acc.	MELD F1.
0.25	69.87	66.92	**66.70**	**64.71**
0.5	**70.79**	**70.93**	63.75	63.22
0.75	69.81	66.24	64.98	61.77
1.0	67.96	65.52	63.03	62.79

Table 9: Hyperparameter sensitivity on λ_{emo}.

λ_{al}	IEMOCAP Acc.	IEMOCAP F1.	MELD Acc.	MELD F1.
0.001	68.45	65.01	63.45	63.22
0.005	68.33	64.19	65.40	63.30
0.01	**70.28**	**70.79**	**66.70**	**64.71**
0.05	68.21	65.11	64.10	62.46

Table 10: Hyperparameter sensitivity on λ_{al}.

5 Conclusion

Building upon a pre-trained text encoder (RoBERTa) and an audio encoder (WavLM or Wav2vec), we introduce a cross-attention mechanism within the text encoder to derive audio-referred text representations, with a modality alignment module to mitigate cross-modal discrepancy. Moreover, we propose an emotion-wheel-guided supervised contrastive loss based on Russell's circumplex model, which accounts for proximity between emotion labels and facilitates emotion-structure-aware representation learning. Extensive experiments on IEMOCAP and MELD demonstrate the superiority of our approach.

Limitation

Although our proposed framework effectively fuses audio and text modalities, it does not incorporate visual modality, which is another effective cue for emotion recognition. As future work, we aim to extend our structure to include the visual modality. Secondly, the performance varies depending on the selection of audio pre-trained models. This gap stems from differences in how well each model's representations align with emotional categories.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(**RS-2024-00345398**). In addition, we would like to acknowledge that this work benefited from mentorship provided through the SK Telecom AI Fellowship Program.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Feiyu Chen, Jie Shao, Shuyuan Zhu, and Heng Tao Shen. 2023. Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10761–10770.

Michal Ephratt. 2011. Linguistic, paralinguistic and extralinguistic speech and silence. *Journal of pragmatics*, 43(9):2286–2307.

Lisa Feldman Barrett and James A Russell. 1998. Independence and bipolarity in the structure of current affect. *Journal of personality and social psychology*, 74(4):967.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5666–5675.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in neural information processing systems*, pages 18661–18673.

Bryan Li, Dimitrios Dimitriadis, and Andreas Stolcke. 2019. Acoustic and lexical sentiment analysis for customer service calls. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5876–5880.

Hang Li, Wenbiao Ding, Yu Kang, Tianqiao Liu, Zhongqin Wu, and Zitao Liu. 2021a. Ctal: Pretraining cross-modal transformer for audio-and-language representations. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 3966–3977.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021b. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in neural information processing systems*, pages 9694–9705.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. 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 3469–3483.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: an attentive rnn for emotion detection in conversations. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 527–536.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.

Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3):715–734.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. 2013. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.

Kailai Yang, Tianlin Zhang, Hassan Alhuzali, and Sophia Ananiadou. 2023. Cluster-level contrastive learning for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, 14(4):3269–3280.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320.

Tao Zhang and Zhenhua Tan. 2025. ECERC: Evidence-cause attention network for multi-modal emotion recognition in conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2064–2077.

Botao Zhao, Xiaoyang Qu, Zuheng Kang, Junqing Peng, Jing Xiao, and Jianzong Wang. 2025. Accon: Angle-compensated contrastive regularizer for deep regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22750–22758.

A Proof of Adjusted Cosine Similarity

Given Eq. 5, we have

$$\cos(\tilde{\theta}_{i,m}) = \cos(\hat{\theta}_{i,m} + \phi). \tag{A.1}$$

Based on the trigonometric identities, we can derive:

$$\begin{aligned}\cos(\tilde{\theta}_{i,m}) &= \cos(\hat{\theta}_{i,m})\cos(\phi) - \sin(\hat{\theta}_{i,m})\sin(\phi) \\ &= \cos(\hat{\theta}_{i,m})\cos(\phi) \\ &\quad - |\sin(\phi)|\sqrt{1-\cos^2(\hat{\theta}_{i,m})}\,.\end{aligned} \tag{A.2}$$

As $\cos(\hat{\theta}_{i,m})$ is equal to the inner product between the embeddings:

$$\cos(\hat{\theta}_{i,m}) = \mathbf{Z}_i^\top \mathbf{Z}_m, \tag{A.3}$$

we obtain the adjusted cosine similarity:

$$\begin{aligned}\cos(\tilde{\theta}_{i,m}) &= (\mathbf{Z}_i^\top \mathbf{Z}_m)\cos(\phi) \\ &\quad - |\sin(\phi)|\sqrt{1-(\mathbf{Z}_i^\top \mathbf{Z}_m)^2+\epsilon}\,,\end{aligned} \tag{A.4}$$

where ϵ is a small constant added for numerical stability.

B Modality-wise Analysis

Table B.1 presents a performance comparison between EMART and ECERC, the previous SOTA method, across different modality combinations. T, A, and V represent text, audio, and visual modalities, respectively. EMART achieves superior or comparable performance to ECERC even when using only audio and text modalities, while ECERC relies on additional video information to reach its best performance. Notably, ECERC suffers from performance degradation when audio is incorporated with text on MELD. In contrast, our method consistently benefits from audio incorporation on both datasets.

Method	Modality	IEMOCAP		MELD	
		Acc.	F1.	Acc.	F1.
ECERC	T	59.70	60.03	66.67	65.47
	T+A	68.15	68.36	65.75	64.48
	T+A+V	69.87	70.10	66.09	**65.15**
OURS	T	64.70	64.54	64.75	64.27
	T+A	**70.79**	**70.93**	**66.70**	64.71

Table B.1: Modality-wise Performance Comparison

For further analysis, we present modality-wise case studies in Fig. B.1. These results demonstrate the effectiveness of our audio-referred text representation strategy, which treats the text as the primary modality and utilizes audio as a complementary reference to mitigate cross-modal interference.

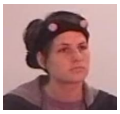
Instance	Label
It's becoming a habit with you... 	**Ground Truth: neutral** ECERC (T): angry ✗ ECERC (T+A): angry ✗ **ECERC (T+A+V): neutral** ✓ Ours (T): angry ✗ **Ours (T+A): neutral** ✓
No, I know me either.... 	**Ground Truth: sad** ECERC (T): frustrated ✗ ECERC (T+A): neutral ✗ **ECERC (T+A+V): sad** ✓ Ours (T): frustrated ✗ **Ours (T+A): sad** ✓
Because she knows what I know - she's faithful as a rock and my worst moments I think of her waiting... 	**Ground Truth: frustrated** ECERC (T): ang ✗ ECERC (T+A): ang ✗ **ECERC (T+A+V): frustrated** ✓ Ours (T): ang ✗ **Ours (T+A): frustrated** ✓

Figure B.1: Case studies of modality-wise predictions. Our method correctly predicts using T+A only, whereas ECERC requires video input (T+A+V).

Models	Emotion Categories of **IEMOCAP**						**Overall (weighted)**	
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Acc.	F1.
DialogueRNN	29.17±1.84	71.97±7.98	57.64±2.62	61.57±2.45	72.35±5.12	62.38±1.69	61.51±0.99	61.31±0.83
DialogueGCN	27.32±2.80	58.50±4.59	53.30±17.45	52.55±3.45	77.70±3.30	61.59±14.20	58.14±1.23	57.47±0.97
MMGCN	28.01±5.40	74.70±1.08	68.23±0.26	68.63±0.90	76.37±1.39	60.81±1.49	65.43±0.55	64.91±0.76
MM-DFN	30.33±10.42	78.50±5.20	65.45±1.96	70.20±1.23	68.68±4.26	69.82±2.40	66.42±1.69	66.05±1.92
M3Net	62.50±7.51	75.37±2.46	71.79±5.62	68.24±1.18	71.01±2.89	59.67±5.72	68.15±0.81	68.36±0.63
CTAL-base	37.27±6.45	60.00±2.48	65.10±3.16	56.47±5.97	44.81±9.97	59.14±0.55	55.82±1.56	55.86±1.53
CTAL-large	34.72±5.56	53.20±3.96	57.20±5.21	62.35±2.70	50.39±5.70	55.67±1.68	53.87±0.78	54.00±0.60
ECERC	59.03±5.68	75.92±2.86	75.09±7.55	69.02±4.75	65.00±4.37	63.43±0.40	68.56±0.51	68.70±0.55
EMART (WavLM)	46.99±11.71	76.73±4.26	76.74±5.56	60.39±3.02	72.80±5.76	70.78±5.39	**70.28±0.49**	**70.29±0.56**

	Emotion Categories of **MELD**							**Overall (weighted)**	
	Neutral	Surprise	Fear	Sad	Joy	Disgust	Angry	Acc.	F1.
DialogueRNN	79.30±1.86	45.91±3.39	0.00±0.00	24.20±4.00	51.08±1.90	0.00±0.00	49.85±7.21	59.49±0.42	57.63±0.08
DialogueGCN	78.53±1.09	34.28±11.51	0.00±0.00	6.41±1.82	40.63±6.75	0.00±0.00	52.17±1.45	55.15±1.41	52.34±1.43
MMGCN	80.95±4.02	45.67±0.14	0.00±0.00	15.55±6.94	52.82±8.91	0.00±0.00	46.38±15.15	59.37±1.52	56.68±0.73
MM-DFN	80.63±0.69	49.82±3.11	0.00±0.00	20.03±1.47	50.17±1.52	0.00±0.00	46.38±1.90	59.62±0.38	57.41±0.41
M3Net	83.55±0.77	60.14±0.71	12.00±4.00	25.48±3.15	64.34±1.01	21.57±4.73	48.98±5.92	65.89±0.12	64.34±0.23
CTAL-base	76.91±4.63	49.47±1.43	8.00±5.29	16.67±2.65	36.32±4.76	9.80±5.17	30.34±6.69	53.68±1.06	51.23±0.34
CTAL-large	81.5±4.00	49.11±4.94	4.67±4.16	15.71±6.40	35.32±4.09	0.00±0.00	18.65±4.66	53.75±0.44	49.39±1.54
ECERC	82.64±0.42	59.9±1.08	12.00±0.00	30.93±1.11	63.93±1.32	22.06±3.89	46.67±1.81	65.50±0.07	64.28±0.02
EMART (Wav2vec)	86.31±1.9	52.31±5.24	9.33±4.16	28.04±4.57	59.79±0.58	6.37±2.25	56.81±0.29	**66.46±0.30**	**64.56±0.13**

Table B.2: Mean and standard deviation over three random seeds on IEMOCAP and MELD.

C Evaluation Across Random Seeds

Table B.2 reports the mean and standard deviation of the main experimental results across three random seeds, using the best-performing audio encoder (WavLM for IEMOCAP and Wav2Vec2 for MELD). All experiments are reproduced using their original implementations.