Multimodal Anxiety Discorder Detection Based on Clinical Interview

Anonymous ACL submission

Abstract

With the rapid development of artificial intelligence, multimodal methods have received increasing attention in the field of mental health disorder detection. Most of the existing research focuses on depression and schizophrenia, but there are relatively few studies on anxiety disorders. To further explore the clinical applicability of multimodal learning in anxiety disorder detection, we propose Multimodal Anxiety Detection via Clinical Interviews (MADCI), a framework designed to automatically identify anxiety disorders from real-world patient-doctor interview data. MADCI comprises three main components: modality-specific feature extractors, a hierarchical cross-modal attention fusion module, and a residual-enhanced multilayer perceptron classifier. In particular, the hierarchical cross-modal attention fusion module captures semantic correlations and complementary information across modalities by integrating cross-modal interactions at multiple levels, thereby enhancing the robustness and discriminative capacity of the fused representations. The validity of MADCI was verified on the MMDA dataset, and its performance was significantly better than that of the current state-of-the-art multimodal models.

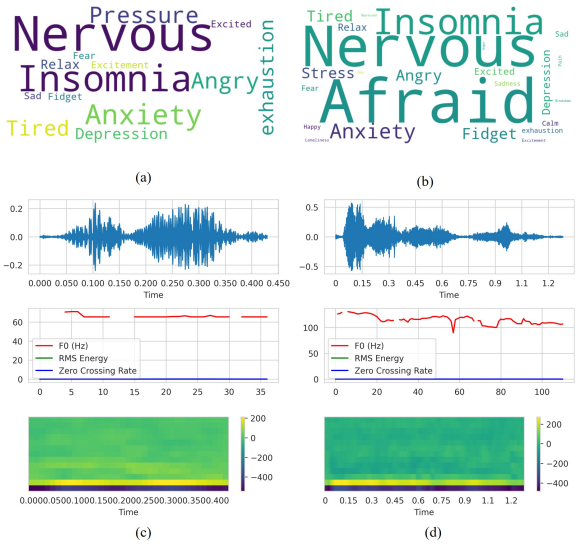

Figure 1: Comparison of emotional word clouds and audio features between normal subjects and anxiety patients. (a) Emotional word cloud generated from interview texts of normal subjects. (b) Emotional word cloud generated from interview texts of anxiety patients. (c) Visualization of audio features from normal subjects, including waveform, F0, RMS energy, zero crossing rate, and spectrogram. (d) Visualization of audio features from anxiety patients, highlighting speech signal energy and frequency differences.

1 Introduction

Anxiety disorder is a prevalent emotional mental illness characterized primarily by persistent tension, worry, and fear, which severely impairs patients' daily quality of life (Sarmiento and Lau, 2020). According to statistics from the World Health Organization (WHO), approximately 3.6% of the global population suffers from anxiety disorders, with the prevalence showing an increasing trend annually (Organization, 2017). Anxiety not only diminishes patients' quality of life but may also lead to comorbidities such as depression and cardiovascular diseases (Bandelow and Michaelis, 2015). Consequently, accurate and objective detection and classification of anxiety patients are of significant importance for early intervention, precise treatment, and mental health management.

Currently, the diagnosis of anxiety disorders primarily relies on psychological questionnaires, clinical interviews, and self-assessment scales (Association, 2013). However, these methods have limitations such as strong subjectivity, low efficiency, susceptibility to situational influences, and a lack of real-time monitoring capabilities. In recent years, with the continuous advancement of multimodal technologies, it has become possible to achieve more efficient and objective recognition of emotional disorders by automatically analyzing patients' linguistic expressions, acoustic features,

facial expressions, and behavioral data during interviews (Abdullah et al., 2021) using computational methods. Multimodal learning not only captures information from different modalities but also extracts complementary interactions between them, thereby enhancing the model's ability to understand the complex psychological states of individuals with anxiety (Tang et al., 2017).

Fig. 1 and previous studies (Ekman, 1992) indicate that patients with anxiety differ from others in terms of text, language and vision. Relying solely on a single model may lead to incomplete assessment, be susceptible to noise, and be overly subjective. This problem is particularly evident in complex clinical settings, where patients may intentionally or unintentionally hide symptoms in one form and manifest them in others. To address the shortcomings of single-modal methods, researchers have increasingly turned to multimodal fusion techniques such as early and late fusion, which integrate diverse signals to enable more accurate and holistic mental health assessments.

Early fusion (Baltrušaitis et al., 2019) strategies typically operated at the feature level, where low-level features from different patterns were concatenated and fed into a unified classifier for joint learning. In contrast, the late-stage fusion (Atrey et al., 2010) strategy models each modality independently and combines their predictions at the decision-making level. However, these methods still face two major challenges: (1) The semantic gap between heterogeneous modalities hinders simple join or summation operations (Ramachandram and Taylor, 2017); (2) The dynamic dependency relationship between patterns changes with the variation of context information, making the fixed fusion strategy not optimal.

To solve these problems and make full use of the information of different modalities, we propose a multimodal fusion anxiety detection model, called Multimodal Anxiety Detection through Clinical Interviews (MADCI). This model uses a modal-specific feature extractor to capture the feature information of each modality, and uses a hierarchical cross-modal attention fusion module to integrate the features of different modalities. Finally, the fused features are classified through the residue-enhanced multi-layer perceptron to determine whether the patient shows anxiety symptoms.

The main contributions of this work are summarized as follows:

- To improve the performance of the anxiety diagnosis task, MADCI designs specific feature encoders for each modality, which can extract the features suitable for the anxiety diagnosis task more effectively.

- MADCI adopts a Hierarchical Cross-Modal Attention Fusion mechanism, enabling the model to focus on the most significant features within and between modalities, hereby promoting the information interaction among text, audio and video and improving the classification performance.

- MADCI can assist clinicians in rapidly assessing patients' anxiety status, thereby shortening initial screening time, improving diagnostic efficiency, and alleviating the shortage of mental health resources.

2 Related Work

In recent years, multimodal sentiment analysis and psychological disorder recognition have emerged as critical research directions in AI-driven mental health studies. Existing research primarily focuses on the following aspects.

2.1 Machine Learning Apporaches for Anxiety Detection

Early research on anxiety disorder detection primarily relied on manual assessment tools and traditional machine learning algorithms (Low et al., 2010). To diagnose anxiety in patients, at least 20 minutes are required for an interview with the patient, during which the patient must correctly understand the questionnaire content and complete the questionnaire (Arif et al., 2020). This mainly depends on the patient's subjective feedback, making it difficult to achieve efficient and objective anxiety screening. To overcome the limitations of manual assessment, researchers have focused on using machine learning algorithms to assist clinicians in diagnosing anxiety disorders, as machine learning has been applied in various fields.

For example, text-based methods use Bag of Words models (Qader et al., 2019), TF-IDF (Ramos, 2003), or LIWC (Tausczik and Pennebaker, 2010) to extract linguistic features, followed by classifiers like SVM (Wang and Hu, 2005), Logistic Regression (Peng et al., 2002), or Random Forest (Rigatti, 2017) to perform binary anxiety prediction. Niva et al. analyzed the blink

data of 44 participants aged 18-30 using machine learning techniques, achieving detection rates ranging from 88% to 94% with ten-fold cross-validation (Das et al., 2025). Li et al. developed a model using MRI to quantify EPVS markers and machine learning algorithms to assess the severity of anxiety and depression symptoms in patients who have used mobile phones for extended periods (Li et al., 2025). Abdulrahman et al. used machine learning models to analyze the distribution characteristics of various physiological signals (Alkurdi et al., 2025). Ancillon et al. reviewed research on anxiety detection using biosignals combined with machine learning methods, systematically analyzing the strengths, weaknesses, and challenges of different signal types, feature extraction methods, and classification models (Ancillon et al., 2022). Bhatnagar et al. (Bhatnagar et al., 2023) collected questionnaire data from university students and used machine learning algorithms to detect and classify the anxiety levels of students.

While these models demonstrate potential in anxiety detection applications, they largely rely on handcrafted features, lack the ability to model nonlinear relationships and cross-modal dependencies, and are limited in their generalization ability in clinical settings.

2.2 Deep Learning Approaches for Anxiety Detection

Multimodal anxiety emotion recognition, by integrating multi-source behavioral signals and depicting an individual's psychological state from multiple dimensions, has become one of the key directions in the current research on intelligent recognition of mental disorders. Diep et al. (Diep et al., 2022) collected the speech and text data of the subjects in the self-management speech task, and extracted deep learning features and manual features from them. Among them, the F1 score of anxiety detection increased by 3% compared with the model that only used manual features. The Multimodal Transformer model proposed by Tsai et al. (Tsai et al., 2019) introduces a cross-modal attention mechanism, which can explicitly model interaction dependencies and temporal dynamics among different modalities, providing a powerful modeling ability for the recognition of complex emotional states.

Although the above-mentioned deep learning methods have certain advantages in specific scenarios, they generally have problems such as insufficient information volume, weak anti-interference ability and poor generalization. Especially in the context of complex and changeable emotional states such as anxiety, relying solely on a certain modal often makes it difficult to capture complete emotional characteristics and meet the requirements of high-precision recognition.

3 Multimodal Anxiety Detection via Clinical Interviews

To effectively identify anxiety states in patients during clinical interview interactions, we propose a multimodal anxiety diagnosis model based on clinical interviews. The overall architecture is illustrated in Fig. 2.

3.1 Modality-Specific Feature Extractors

3.1.1 Text Feature Extraction

In multimodal mental health analysis tasks, the text modality often carries rich semantic information. This is particularly true in doctor-patient dialogue scenarios, where patients' linguistic expression, emotional tendencies, and language structure can all reflect their mental state (Cambria and White, 2014). Therefore, the effective extraction of textual features is crucial for the detection of anxiety disorders. In this work, we employ BERT (Devlin et al., 2019) as the encoder for the text modality, leveraging its bidirectional Transformer architecture (Han et al., 2022) for deep semantic modeling of context.

Specifically, we isolate and encode only the patient's spoken content using the pretrained bert-large-uncased model (Yu et al., 2022). Given an input sentence consisting of n words $S = \{w_1, w_2, \ldots, w_n\}$, the corresponding dialogue is tokenized using the WordPiece tokenizer into a sequence $T = \{[\text{CLS}], t_1, t_2, \ldots, t_k, [\text{SEP}]\}$.

This token sequence is then mapped into embedding vectors $E \in \mathbb{R}^{k \times d}$, where $d = 1024$ denotes the hidden dimension. The embeddings are fed into a 24-layer Transformer encoder to obtain contextualized representations for each token. The semantic representation of the entire sequence is given by the output corresponding to the [CLS] token, formulated as $T_{\text{raw}} = \text{BERT}_{\text{large}}(T)[0] \in \mathbb{R}^{1024}$.

To further enhance the non-linear modeling capacity of the extracted features, we design a residual fully connected block to transform T_{raw}. This block contains a ReLU-activated dense layer and a shortcut path aligned in dimension. The two

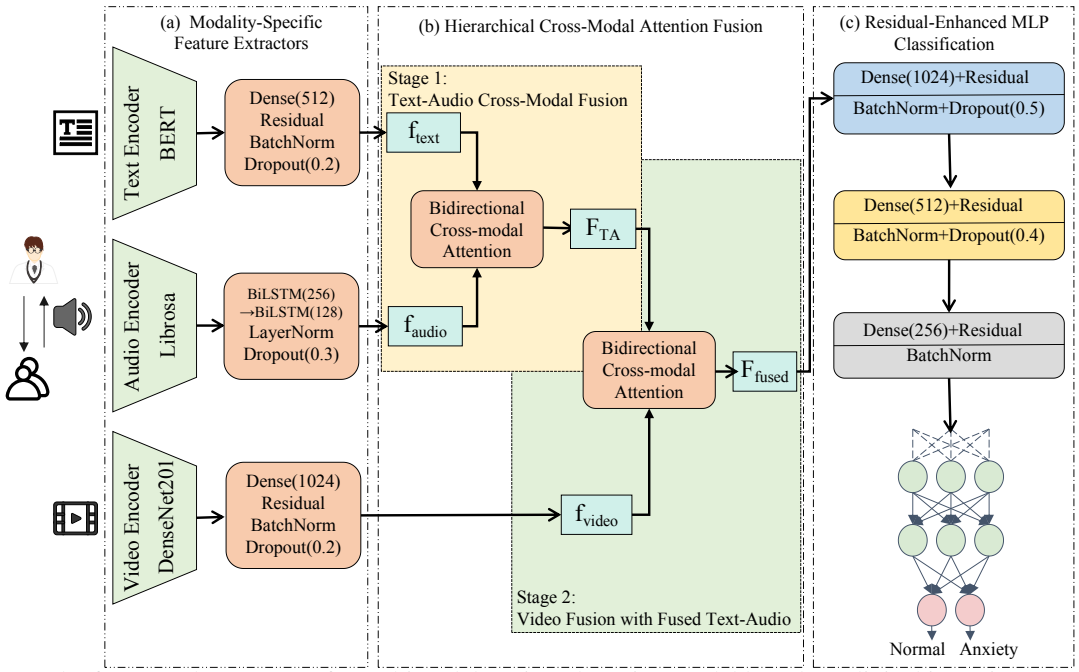

Figure 2: The framework structure of a multimodal anxiety detection model based on clinical interviews (MADCI). (a) Modality-Specific Feature Extractors, which independently encode text, audio, and video inputs; (b) a Hierarchical Cross-Modal Attention Fusion Module, designed to capture inter-modal interactions at multiple semantic levels; and (c) a Residual-Enhanced MLP Classifier, which integrates the fused features for final prediction.

branches are combined as the final output. The main transformation branch is defined as Eq. (1):

$$T' = \text{ReLU}(W_1 T_{\text{raw}} + b_1), W_1 \in \mathbb{R}^{512\times1024} \quad (1)$$

If the input and output dimensions are inconsistent, a linear projection is applied to the shortcut path:

$$T_{\text{shortcut}} = W_2 T_{\text{raw}} + b_2, W_2 \in \mathbb{R}^{512\times1024} \quad (2)$$

Otherwise, the shortcut is directly taken as $T_{\text{shortcut}} = T_{\text{raw}}$. The residual output is then obtained by combining the main and shortcut branches $T_{\text{res}} = T' + T_{\text{shortcut}}$. To improve the robustness and generalization ability of the model, we apply Batch normalization and Conditional Dropout to T_{res}. Specifically, the transformation is formulated as $f_{\text{text}} = \text{Dropout}(\text{BN}(T_{\text{res}}), p = 0.2)$, where BN denotes Batch Normalization. The resulting feature f_{text} serves as the intermediate semantic representation of the text modality and is fed into the subsequent multimodal fusion module.

3.1.2 Audio Feature Extraction

To capture rich acoustic information for downstream anxiety detection, we focus on the speech modality derived from clinical interview sessions. Based on the timestamp annotations in the doctor-patient dialogue transcripts, we segment the original audio recordings and retain only the segments corresponding to the patient's speech. This ensures temporal alignment with other modalities. And use the librosa library to extract low-level and high-level audio features from the original speech, and combine time-frequency domain features with manually designed features to form complementary acoustic representations.

Finally, an 185-dimensional audio feature vector was extracted, including MFCC (20-d) (Zheng et al., 2001), Chroma (24-d), Mel-Spectrogram (128-d), and 13 low-level descriptors (e.g., zero-crossing rate, energy, spectral centroid). All features were standardized via z-score normalization. Given an audio signal $x(t)$, we compute its Short-Time Fourier Transform (Benesty et al., 2011), apply the Discrete Cosine Transform (Khayam, 2003), and obtain average MFCCs over T frames. Chroma features are extracted using a 24-bin constant-Q transform, enhancing pitch sensitivity. The final feature vector is defined as $A = [\text{MFCC}_{1:20}, \text{Chroma}, \text{Mel}, \text{Contrast}, \text{Tonnetz}]$.

To incorporate acoustic features into the model, we adopt a lightweight BiLSTM structure. Since A is non-temporal, we reshape it using $\text{ExpandDims}(A)$ to simulate sequential input. The encoded representation is obtained as: $h_1 = \text{BiLSTM}_{256}(\text{ExpandDims}(A))$, $h_2 = \text{BiLSTM}_{128}(h_1)$, followed by layer normalization to obtain the final audio feature $f_{\text{audio}} =$

LayerNorm(h_2), where BiLSTM$_N$ denotes a bidirectional LSTM with N hidden units per direction. LayerNorm is applied to stabilize training and enhance convergence.

3.1.3 Video Feature Extraction

The video modality serves as a critical source of emotional cues, offering valuable visual information such as facial expressions, body posture, and eye movement, which are essential for evaluating a patient's mental state (Canal et al., 2022). In this study, we adopt DenseNet201, pre-trained on ImageNet, as the backbone network to extract deep visual representations from clinical interview videos. This network captures both local texture features (e.g., facial muscle movements, wrinkles, skin tone variation) and global visual semantics (e.g., head pose, gaze aversion), providing insights into psychological health via spatiotemporal cues.

To ensure temporal alignment with other modalities, we segment the video according to the timestamp annotations in the transcript JSON files, extracting only the segments during which the patient is speaking. For each segment, we uniformly sample a fixed number of keyframes and resize them to 224×224 pixels. All frames are normalized before being fed into DenseNet201.

For each sampled frame F_i, a 1920-dimensional deep semantic representation is extracted using the DenseNet201 model, formulated as $\mathbf{x}_{\text{frame}}^{(i)} = \text{DenseNet201}(F_i)$, where $\mathbf{x}_{\text{frame}}^{(i)} \in \mathbb{R}^{1920}$. To summarize the temporal dynamics of a segment, we apply global average pooling across all M frames, resulting in the aggregated feature vector $\mathbf{V} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{x}_{\text{frame}}^{(i)}$.

To enhance feature expressiveness and mitigate vanishing gradients, we design a Residual Dense Block. The global feature vector $\mathbf{V}$ is projected into a lower-dimensional latent space via a fully connected layer as $\mathbf{V}' = \text{ReLU}(\mathbf{W}_v\mathbf{V} + \mathbf{b}_v)$, where $\mathbf{V}' \in \mathbb{R}^{1024}$. To ensure compatibility for residual connections, the original input $\mathbf{V}$ is projected into the same latent space through a linear transformation, formulated as $\mathbf{V}_{\text{proj}} = \mathbf{W}_p\mathbf{V} + \mathbf{b}_p$, where $\mathbf{W}_p \in \mathbb{R}^{1024\times1920}$. The final residual feature representation is then computed by element-wise addition $\mathbf{V}_{\text{res}} = \mathbf{V}' + \mathbf{V}_{\text{proj}}$. To improve stability and generalization, we apply batch normalization and dropout:

$$\text{BN}(\mathbf{V}_{\text{res}}) = \gamma \cdot \frac{\mathbf{V}_{\text{res}} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{3}$$

$$\mathbf{f}_{\text{video}} = \text{Dropout}(\text{BN}(\mathbf{V}_{\text{res}}), p = 0.2) \tag{4}$$

where μ and σ^2 denote the batch-wise mean and variance, γ and β are learnable parameters, and ϵ is a small constant for numerical stability. The resulting feature vector $\mathbf{f}_{\text{video}}$ is used as the input to the multimodal fusion network.

3.2 Hierarchical Cross-Modal Attention Fusion

We propose a Hierarchical Cross-Modal Attention Fusion (HCAF) module. The core innovation lies in a two-stage attention mechanism that performs fine-grained alignment followed by global integration. Specifically, we introduce a local-to-global attention design, where local (fine-grained) interactions between two modalities are first captured, and then aggregated through global multi-modal integration. As illustrated in Fig. 3.

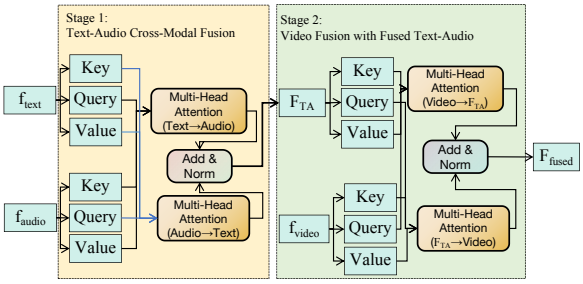

Figure 3: Hierarchical Cross-Modal Attention Fusion Framework for Multimodal Anxiety Detection. The diagram illustrates a hierarchical fusion strategy where textual and audio features are first integrated via bidirectional cross-modal attention, followed by fusion with video features. Each fusion step employs multi-head attention, global pooling, and residual connections to preserve and enhance multimodal information representation.

3.2.1 Modality Alignment Projection Layer

To address the inherent heterogeneity of multimodal feature spaces, we design a modality-specific projection network to map each modality into a unified latent space while preserving modality characteristics. Given the extracted modality-specific features $\mathbf{f}_{\text{text}} \in \mathbb{R}^{d_t}$, $\mathbf{f}_{\text{audio}} \in \mathbb{R}^{d_a}$, and $\mathbf{f}_{\text{video}} \in \mathbb{R}^{d_v}$, we project them into a common feature space of dimension $d = 256$ via independent fully connected layers, formulated as $\mathbf{T}' = \mathbf{W}_t\mathbf{f}_{\text{text}} + \mathbf{b}_t$, where $\mathbf{W}_t \in \mathbb{R}^{d\times d_t}$ for the text modality, $\mathbf{A}' = \mathbf{W}_a\mathbf{f}_{\text{audio}} + \mathbf{b}_a$, where $\mathbf{W}_a \in \mathbb{R}^{d\times d_a}$ for the audio modality, and $\mathbf{V}' =$

$\mathbf{W}_v\mathbf{f}_{\text{video}} + \mathbf{b}_v$, where $\mathbf{W}_v \in \mathbb{R}^{d\times d_v}$ for the video modality.

All projection matrices $\mathbf{W}_t$, $\mathbf{W}_a$, and $\mathbf{W}_v$ are initialized using orthogonal initialization to mitigate modality dominance during training. To further promote modality balance and adaptivity, we introduce a learnable scaling factor α_m for each modality $m \in \{\text{text}, \text{audio}, \text{video}\}$, which dynamically adjusts the contribution of each modality based on its projected representation, defined as $\alpha_m = \sigma(\mathbf{v}_m^\top \mathbf{M}')$ with $\mathbf{v}_m \in \mathbb{R}^d$. where $\mathbf{M}' \in \{\mathbf{T}', \mathbf{A}', \mathbf{V}'\}$ corresponds to the projected feature of modality m, and $\sigma(\cdot)$ denotes the sigmoid activation function to ensure $\alpha_m \in (0, 1)$. The adaptively scaled features are then computed as:

$$\tilde{\mathbf{T}} = \alpha_{\text{text}}\cdot\mathbf{T}', \quad \tilde{\mathbf{A}} = \alpha_{\text{audio}}\cdot\mathbf{A}', \quad \tilde{\mathbf{V}} = \alpha_{\text{video}}\cdot\mathbf{V}' \tag{5}$$

This adaptive mechanism enables the model to automatically calibrate the relative importance of each modality under different contexts, while maintaining compatibility for downstream shared attention-based fusion mechanisms.

3.2.2 Bidirectional Cross-Modal Attention Layer

To enable fine-grained interactions across modalities, we propose a hierarchical progressive fusion strategy based on a two-stage cross-modal attention mechanism. This strategy builds upon standard multi-head attention and enhances multimodal representation through step-wise fusion.

Stage 1: Text-Audio Cross-Modal Fusion. Given the modality-specific projected features $\tilde{\mathbf{T}}, \tilde{\mathbf{A}} \in \mathbb{R}^d$, we first perform bi-directional attention between the text and audio modalities using a multi-head attention mechanism with $M = 4$ heads. The general formulation of cross-attention is defined as $\text{CrossAttn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}$, where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{d\times M}$ are computed from the concatenated features $[\tilde{\mathbf{T}}; \tilde{\mathbf{A}}]$ as $\mathbf{Q} = \mathbf{W}_q^\top[\tilde{\mathbf{T}}; \tilde{\mathbf{A}}]$, $\mathbf{K} = \mathbf{W}_k^\top[\tilde{\mathbf{T}}; \tilde{\mathbf{A}}]$, and $\mathbf{V} = \mathbf{W}_v^\top[\tilde{\mathbf{T}}; \tilde{\mathbf{A}}]$, where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{2d\times d_h}$ are learnable parameters and d_h denotes the head dimension. The attention outputs are then processed with a residual connection and layer normalization to stabilize training and preserve original information:

$$\tilde{\mathbf{T}}' = \tilde{\mathbf{T}} + \text{LayerNorm}\left(\text{CrossAttn}_T(\tilde{\mathbf{A}}, \tilde{\mathbf{T}})\right) \tag{6}$$

$$\tilde{\mathbf{A}}' = \tilde{\mathbf{A}} + \text{LayerNorm}\left(\text{CrossAttn}_A(\tilde{\mathbf{T}}, \tilde{\mathbf{A}})\right) \tag{7}$$

To obtain a compact representation, we apply global average pooling to the attended features $\tilde{\mathbf{T}}'$ and $\tilde{\mathbf{A}}'$, and then concatenate the pooled results to form the fused text-audio representation $\mathbf{F}_{\text{TA}} = \text{AvgPool}(\tilde{\mathbf{T}}') \,\|\, \text{AvgPool}(\tilde{\mathbf{A}}') \in \mathbb{R}^{2d}$, where $\text{AvgPool}(\cdot)$ denotes global average pooling and $\|$ denotes vector concatenation.

Stage 2: Video Fusion with Fused Text-Audio. In the second stage, we hierarchically incorporate the video modality by fusing $\mathbf{F}_{\text{TA}}$ with the projected video features $\tilde{\mathbf{V}} \in \mathbb{R}^d$. Following the same cross-attention formulation as in Stage 1, we perform bi-directional attention and residual fusion, resulting in the fused representation $\mathbf{F}_{\text{TAV}} = \text{CrossAttnFusion}(\mathbf{F}_{\text{TA}}, \tilde{\mathbf{V}})$.

3.3 Residual-Enhanced Multi-Layer Perceptron Classification Module

To further enhance the discriminative power of the model and improve the nonlinear representation of the fused features, we employ a residual multilayer perceptron (ResMLP) as the final classifier. Specifically, the fused feature $\mathbf{F}_{\text{fused}}$ is passed through a stack of three residual fully connected (FC) blocks. Each block is formulated as$\mathbf{z} = \text{ReLU}(\mathbf{W}\mathbf{h}_{\text{in}} + \mathbf{b})$, $\mathbf{h}_{\text{out}} = \mathbf{z} + \text{Proj}(\mathbf{h}_{\text{in}})$, where $\mathbf{W}$ and $\mathbf{b}$ denote the learnable weights and biases of the current layer. The projection function $\text{Proj}(\cdot)$ is defined as Eq. (8):

$$\text{Proj}(\mathbf{h}_{\text{in}}) = \begin{cases} \mathbf{h}_{\text{in}}, & \text{if } d_{\text{in}} = d_{\text{out}} \\ \mathbf{W}_{\text{proj}}\mathbf{h}_{\text{in}}, & \text{otherwise} \end{cases} \tag{8}$$

where d_{in} and d_{out} represent the input and output feature dimensions, respectively.

Residual connections help mitigate the vanishing gradient problem, accelerate convergence, and improve generalization by preserving feature propagation through identity mapping. The complete ResMLP pipeline is summarized as follows:

$$\mathbf{h}_1 = \text{Dropout}\left(\text{BN}\left(\text{ResBlock}\left(\mathbf{F}_{\text{fused}}, 1024\right)\right),\ p = 0.5\right) \tag{9}$$

$$\mathbf{h}_2 = \text{Dropout}\left(\text{BN}\left(\text{ResBlock}\left(\mathbf{h}_1, 512\right)\right),\ p = 0.4\right) \tag{10}$$

$$\mathbf{h}_3 = \text{BN}\left(\text{ResBlock}\left(\mathbf{h}_2, 256\right)\right) \tag{11}$$

where $\text{BN}(\cdot)$ denotes batch normalization, and $\text{Dropout}(\cdot, p)$ denotes dropout with probability p. Finally, the output $\mathbf{h}_3$ is mapped to the binary classification probability $\hat{y} \in [0, 1]$ through a sigmoid-activated linear transformation, formulated as $\hat{y} = \sigma(\mathbf{W}^\top\mathbf{h}_3 + b)$, where $\sigma(\cdot)$ denotes the sigmoid

function. As this task is binary classification, we adopt the binary cross-entropy loss as the optimization objective:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{12}$$

$y_i \in \{0, 1\}$ denotes the ground truth label, and $\hat{y}_i$ is the predicted probability for the i-th sample. The model is optimized using the Adam optimizer.

4 Experiments

In this study, we conducted the validation of the experimental validity on the MMDA (Jiang et al., 2022) clinical interview dataset. For the detailed introduction of the dataset, please refer to the appendix.

4.1 Comparison with State-of-the-Art Models

To further validate the effectiveness of the proposed MADCI in multimodal fusion scenarios, we compare it with several representative multimodal learning baselines on the preprocessed MMDA dataset. The comparison includes MuIT (Bhattacharjee et al., 2022), FLAVA (Singh et al., 2022), Data2Vec2.0 (Baevski et al., 2023), and CrossNet. All models utilize three modalities: text, audio, and video. Table 1 summarizes the performance in terms of accuracy, precision, recall, and F1-score. As shown, MADCI consistently outperforms existing methods across all evaluation metrics, demonstrating its superior capacity for multimodal representation learning.

Modality	Method	Pub.	ACC	Precision	Recall	F1-Score	AUC
T+A+V	MuIT	CVPR22	0.802	0.789	0.666	0.779	0.850
	FLAVA	CVPR22	0.792	0.781	0.634	0.759	0.848
	Data2Vec2.0	PMLR23	0.812	0.796	0.601	0.771	0.843
	MADCI	Ours	**0.871**	**0.876**	**0.721**	**0.854**	**0.883**

Table 1: Comparison of model performance on the multimodal classification task. T, A, and V denote the text, audio, and video modalities, respectively.

As shown in Table 1, the proposed MADCI model consistently outperforms all baselines across evaluation metrics. This improvement stems from its multi-head hierarchical attention architecture, which enables fine-grained cross-modal interactions at various abstraction levels, effectively capturing complex inter-modal dependencies and addressing the limitations of traditional fusion approaches. Additionally, modality-specific encoders and residual dense blocks enhance feature representation by preserving deep semantic information and improving fusion quality.

In contrast, baseline models show notable weaknesses. MuIT relies on token-level fusion and lacks dynamic weighting across modalities, reducing adaptability to noisy inputs. FLAVA suffers from low recall and F1 scores, likely due to limited temporal modeling, leading to misclassification near decision boundaries. Data2Vec 2.0 achieves relatively balanced performance but lags behind MADCI, possibly due to its generic self-supervised features and the absence of explicit modality alignment.

4.2 Ablation Experiment

The effectiveness of Hierarchical Cross-Modal Attention Fusion: To investigate the impact of multimodal fusion strategies on model performance and to identify the most effective method for affective disorder recognition, we conduct a comprehensive comparison of several representative fusion approaches. Table 2 summarizes the performance of each fusion strategy on the MMDA dataset for anxiety disorder classification.

Fusion Strategy	ACC	Precision	Recall	F1-score
Early fusion	0.832	0.821	0.712	0.823
Attention fusion	0.851	0.857	0.675	0.826
Cross model fusion	0.822	0.805	0.673	0.806
Gated Attention fusion	0.822	0.808	0.689	0.811
Hierarchical Fusion (Ours)	**0.871**	**0.876**	**0.721**	**0.854**

Table 2: Ablation study on different multimodal fusion strategies. We compare early fusion, attention-based fusion, cross-modal fusion, and gated attention fusion. Our hierarchical fusion method achieves the best overall performance across all metrics.

Early Fusion performs direct concatenation of raw modality features at the input level. While computationally simple and efficient, it is often sensitive to differences in scale and distribution across modalities. Attention-based Fusion introduces modality-specific attention weights, enabling the model to emphasize salient features and suppress irrelevant information. Cross-Modal Fusion seeks to align features across modalities by modeling inter-modal correlations explicitly. Gated Attention Fusion further incorporates gating mechanisms to dynamically modulate the contribution of each modality under different contexts. In contrast, the hierarchical fusion adopted in the MADCI model simulates human cognitive processes by gradually integrating multimodal information across abstract

levels. It achieves the best results in all indicators, highlighting its superior ability to effectively utilize the complementary information of text, audio, and visual patterns, and is conducive to more in-depth modeling of complex cross-modal dependencies.

Configuration	Residual	Dropout	ACC	Precision	Recall	F1-Score
No Residual (1024-512-256)	×	[0.5, 0.4]	0.842	0.836	0.669	0.817
No Dropout (1024-512-256)	✓	None	0.802	0.775	0.595	0.764
Shallow(1024-512)	✓	None	0.832	0.817	0.679	0.815
Deeper (1024-768-512-256)	✓	[0.5]	0.842	0.847	0.653	0.811
Wider(1024-1024-512)	✓	[0.5, 0.4, 0.3]	0.802	0.842	0.545	0.731
Ours(1024-512-256)	✓	**[0.5, 0.4]**	**0.871**	**0.876**	**0.721**	**0.854**

Table 3: Ablation study on key components of the classifier. We examine the effects of residual connections, depth, width, and dropout on multimodal anxiety detection performance. The "Configuration" column specifies the architecture's layer dimensions.

The effectiveness of Residual-Enhanced MLP Classification: We conduct a systematic ablation study on ResMLP architectural components. Specifically, we assess the model performance under different configurations by removing residual connections and dropout, reducing or increasing the network depth (with 2 and 4 hidden layers, respectively), and widening the hidden layers.

As shown in Table 3, the ResMLP module achieves the best overall performance. While residual connections slightly increase model complexity, they effectively mitigate the vanishing gradient problem in deep networks. Dropout helps prevent overfitting by regularizing the network. Shallow networks converge faster but often suffer from limited representational capacity, whereas deeper architectures offer stronger expressiveness at the cost of increased training difficulty. Similarly, wider layers can capture more complex patterns but require more parameters.

To further analyze the anxiety detection performance of the MADCI model in different modes: Tables 4 and 5 present the performance of anxiety detection tasks using different multimodal models in single-modal and dual-modal combinations respectively. In the single-modal text task, FLAVA performs the best, followed by MuIT. In other single-modal and dual-modal anxiety detection tasks, the MADCI we proposed achieved the best performance.

Compared to large multimodal models like MuIT, FLAVA, and Data2Vec 2.0, the proposed MADCI demonstrates superior suitability for anxiety detection. While MuIT offers dynamic cross-modal interaction, it requires substantial computational resources and large annotated datasets, limiting practicality. FLAVA shows stable performance but struggles to capture fine-grained emotional cues. Data2Vec 2.0, though modality-agnostic, relies on teacher-generated pseudo-labels, risking bias and loss of modality-specific details.

Modality	Model	ACC	Precision	Recall	F1-Score
T	MuIT	0.812	0.813	0.660	0.782
	FLAVA	**0.832**	**0.845**	**0.686**	**0.805**
	Data2Vec2.0	0.782	0.612	0.5	0.687
	MADCI (Ours)	0.792	0.768	0.539	0.724
A	MuIT	0.792	0.781	0.634	0.759
	FLAVA	0.822	0.824	**0.679**	**0.797**
	Data2Vec2.0	0.782	0.612	0.5	0.687
	MADCI (Ours)	**0.832**	**0.836**	0.630	0.795
V	MuIT	0.743	0.551	0.5	0.633
	FLAVA	0.762	0.736	0.576	0.712
	Data2Vec2.0	0.218	0.047	0.5	0.078
	MADCI (Ours)	**0.802**	**0.796**	**0.693**	**0.799**

Table 4: Performance comparison of different models on individual modalities (Text, Audio, Video).

Modality	Model	Accuracy	Precision	Recall	F1-Score
T+A	MuIT	0.822	0.817	0.692	0.802
	FLAVA	0.841	0.835	0.743	0.831
	Data2Vec2.0	0.782	0.612	0.5	0.687
	MADCI (Ours)	**0.851**	**0.843**	**0.708**	**0.836**
T+V	MuIT	0.762	0.736	0.576	0.712
	FLAVA	0.752	0.719	0.544	0.684
	Data2Vec2.0	0.812	0.848	0.568	0.750
	MADCI (Ours)	**0.832**	**0.824**	**0.646**	**0.803**
V+A	MuIT	0.743	0.551	0.5	0.633
	FLAVA	0.772	0.753	0.595	0.729
	Data2Vec2.0	0.723	0.732	0.609	0.727
	MADCI (Ours)	**0.852**	**0.847**	**0.692**	**0.832**

Table 5: Performance comparison of different models on bimodal combinations in multimodal anxiety detection.

In contrast, MADCI employs a hierarchical fusion strategy to progressively integrate text, audio, and video features. With cross-modal attention, bidirectional interaction, residual connections, batch normalization, and dropout, it enhances cross-modal exchange and semantic alignment, improving model robustness and generalization in anxiety detection.

5 Conclusions

To advance anxiety detection, this paper proposes MADCI. It encodes emotional and nonverbal signals from semantic, acoustic, and visual modalities, each processed separately and fused through a hierarchical strategy with cross-modal attention, bidirectional interaction, and residual connections. Experiments on the MMDA dataset show MADCI achieves 87.13% accuracy and 88.36% AUC.

Limitations

Despite the promising performance of our model in anxiety detection using multimodal clinical interview data, with competitive results in terms of accuracy and AUC, several limitations remain:

First, the study is based on the MMDA clinical interview dataset, where each subject has complete data across text, audio, and video modalities. However, the overall dataset size is limited, and the data collection is geographically and culturally homogeneous, which may hinder the generalizability of the model. Due to the scarcity of publicly available multimodal anxiety datasets, we trained and evaluated the model on a single dataset without cross-domain or transfer learning validation. Future work will focus on validating the model across diverse datasets to better assess its robustness and effectiveness.

Secondly, although the proposed MADCI framework effectively integrates textual, acoustic, and visual modalities for improved anxiety detection, it currently lacks the capability for fine-grained classification of anxiety severity. Extending the model to support multi-level anxiety assessment remains an important direction for future research.

Finally, while this study validates the model using real clinical interview data, MADCI has not yet been deployed or evaluated in actual clinical settings. Future work will focus on investigating its feasibility and performance in real-time clinical applications to further enhance its practicality and clinical applicability.

Ethical Considerations

This work uses previously collected human data from the MMDA dataset. Please see the paper that introduces this dataset (Jiang et al., 2022) for information about the data collection procedure. The authors foresee no ethical problems arising from the work presented here.

References

S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree. 2021. Multimodal emotion recognition using deep learning. *Journal of Applied Science and Technology Trends*, 2(01):73–79.

Abdulrahman Alkurdi, Maxine He, Jonathan Cerna, Jean Clore, Richard Sowers, Elizabeth T. Hsiao Wecksler, and Manuel E. Hernandez. 2025. Extending anxiety detection from multimodal wearables in controlled conditions to real-world environments. *Sensors*, 25(4):1241–1241.

L. Ancillon, M. Elgendi, and C. Menon. 2022. Machine learning for anxiety detection using biosignals: a review. *Diagnostics*, 12(8):1794.

M. Arif, A. Basri, G. Melibari, and 1 others. 2020. Classification of anxiety disorders using machine learning methods: A literature review. *Insights in Biomedical Research*, 4(1):95–110.

American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. American Psychiatric Association.

Pradeep K Atrey, M Anwar Hossain, Abdul El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379.

Alexey Baevski, Abhinav Babu, W. N. Hsu, and Michael Auli. 2023. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning (ICML)*, pages 1416–1429.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

B. Bandelow and S. Michaelis. 2015. Epidemiology of anxiety disorders in the 21st century. *Dialogues in Clinical Neuroscience*, 17(3):327–335.

Jacob Benesty, Jian Chen, and E. A. Habets. 2011. *Speech Enhancement in the STFT Domain*. Springer Science & Business Media.

S. Bhatnagar, J. Agarwal, and O. R. Sharma. 2023. Detection and classification of anxiety in university students through the application of machine learning. *Procedia Computer Science*, 218:1542–1550.

D. Bhattacharjee, T. Zhang, S. Süsstrunk, and M. Salzmann. 2022. Mult: An end-to-end multitask learning transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12031–12041.

Erik Cambria and Björn White. 2014. Jumping nlp curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine*, 9(2):48–57.

F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski. 2022. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617.

Niva Das, Laxmipriya Moharana, Satyajit Nayak, and Aurobinda Routray. 2025. Emotional blink patterns: A possible biomarker for anxiety detection in a hci framework. *SN Computer Science*, 6(3):277–277.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2019. 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.

B. Diep, M. Stanojevic, and J. Novikova. 2022. Multimodal deep learning system for depression and anxiety detection. *arXiv preprint arXiv:2212.14490*.

P. Ekman. 1992. *Facial expressions of emotion: New findings, new questions*.

Kai Han, Yun Wang, Haoyan Chen, Xiaoyang Chen, Junyu Guo, Zicheng Liu, and Dacheng Tao. 2022. A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708.

Y. Jiang, Z. Zhang, and X. Sun. 2022. Mmda: A multimodal dataset for depression and anxiety detection. In *International Conference on Pattern Recognition*, pages 691–702. Springer Nature Switzerland.

Syed A. Khayam. 2003. The discrete cosine transform (dct): theory and application. *Michigan State University*, 114(1):31.

Li Li, Yalan Wu, Jiaojiao Wu, Bin Li, Rui Hua, Feng Shi, and Yeke Wu. 2025. Mri quantified enlarged perivascular space volumes as imaging biomarkers correlating with severity of anxiety depression in young adults with long-time mobile phone use. *Frontiers in Psychiatry*, 16:1532256–1532256.

L. S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen. 2010. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586.

World Health Organization. 2017. Depression and other common mental disorders: Global health estimates. Geneva: World Health Organization.

C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll. 2002. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14.

W. A. Qader, M. M. Ameen, and B. I. Ahmed. 2019. An overview of bag of words; importance, implementation, applications, and challenges. In *2019 International Engineering Conference (IEC)*, pages 200–204. IEEE.

Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108.

J. Ramos. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, volume 242, pages 29–48.

S. J. Rigatti. 2017. Random forest. *Journal of Insurance Medicine*, 47(1):31–39.

C. Sarmiento and C. Lau. 2020. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*, pages 125–129.

Amanpreet Singh, Rui Hu, Vikas Goswami, Gaël Couairon, William Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650.

H. Tang, W. Liu, W. L. Zheng, and B. L. Lu. 2017. Multimodal emotion recognition using deep neural networks. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part IV 24*, pages 811–819. Springer International Publishing.

Y. R. Tausczik and J. W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, and R. Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558.

H. Wang and D. Hu. 2005. Comparison of svm and ls-svm for regression. In *2005 International conference on neural networks and brain*, volume 1, pages 279–283. IEEE.

Yubin Yu, Yun Wang, Jianqiang Mu, Weizhou Li, Shaoyan Jiao, Zhongming Wang, and Yizhen Zhu. 2022. Chinese mineral named entity recognition based on bert model. *Expert Systems with Applications*, 206:117727.

F. Zheng, G. Zhang, and Z. Song. 2001. Comparison of different implementations of mfcc. *Journal of Computer Science and Technology*, 16:582–589.

M. Zimmerman, J. Martin, H. Clark, P. McGonigal, L. Harris, and C. G. Holst. 2017. Measuring anxiety in depressed patients: a comparison of the hamilton anxiety rating scale and the dsm-5 anxious distress specifier interview. *Journal of psychiatric research*, 93:59–63.

A Appendix

A.1 Supplementary explanation of Fig. 1

Fig. 1 presents the comparison between anxious and non-anxious patients during clinical interviews,

including text emotion word clouds and visualizations of acoustic features. As shown in Fig. 1(b), the texts of anxious patients frequently contain negative emotional words such as "afraid", "nervous", "insomnia", "anxiety", "stress", and "Fidget", indicating a clear tendency toward negative emotional expression. In contrast, Fig. 1(a) shows that non-anxious patients predominantly use more neutral and stable emotional words, such as "relax", "exited", "tired" et. al. In order to further compare the differences in acoustic characteristics between the two types of patients, Fig. 1(c) and 1(d) respectively show the visualization results of speech features when non-anxious patients and anxious patients have conversations with doctors. It can be seen from the figures that there are significant differences in speech features such as MFCC and F0 between the two types of patients. Moreover, previous research has indicated that anxious individuals are more likely to exhibit tense and anxious facial expressions (e.g., frowning, lip-biting) and more frequent micro-expression changes (Ekman, 1992).

A.2 Datasets

The MMDA dataset contains clinical interview data of 501 participants conducted by licensed psychologists, among which 108 are healthy controls and 393 are clinically diagnosed anxiety cases. The dataset comprises de-identified original interview videos, manually transcribed dialogue text, and HAMA (Zimmerman et al., 2017) scores. An overview of the MMDA dataset is provided in Table 6.

Dataset	Modality	Sample Size	Scale	Age (Avg./Min/Max)
MMDA	A, V, T	501 (108:393)	HAMA	40.53 / 13 / 83

Table 6: Overview of MMDA Clinical Anxiety Dataset.

To develop a multimodal anxiety disorder detection framework, we extract audio segments from the original video recordings and construct a comprehensive dataset comprising text, audio, and video modalities. To ensure both the accuracy and efficiency of the clinical anxiety diagnosis model, rigorous data preprocessing was conducted on the textual modality. This includes the removal of extraneous whitespace, special characters, and other textual noise. Based on the annotated time segments indicating the patient's speech within the transcription files, we synchronously extracted the corresponding audio and video segments for subsequent multimodal feature extraction.

A.3 Implementation Details

Model optimization is performed using the Adam optimizer with a learning rate of 1×10^{-4}. We apply gradient clipping, dropout, label smoothing (with a factor of 0.1), and weight decay (0.01) to stabilize training. The binary cross-entropy loss is used for supervision. The dataset is split into training and test sets using an 80/20 stratified split (seed = 42). All experiments are conducted with fixed random seeds to ensure reproducibility.