Mitigating Interviewer Bias in Multimodal Depression Detection: An Approach with Adversarial Learning and Contextual Positional Encoding

Enshi Zhang, Christian Poellabauer
Knight Foundation School of Computing & Information Sciences
Florida International University, Miami, FL 33199, USA
{ezhan004, cpoellab}@fiu.edu

Abstract

Clinical interviews are a standard method for assessing depression. Recent approaches have improved prediction accuracy by focusing on specific questions posed by the interviewer and manually selected question-answer (QA) pairs that target mental health content. However, these methods often neglect the broader conversational context, resulting in limited generalization and reduced robustness, particularly in less structured interviews, which are common in real-world clinical settings. In this work, we develop a multimodal dialogue-level transformer that captures the dynamics of dialogue within each interview by using a combination of sequential positional embedding and question context vectors. In addition to the depression prediction branch, we build an adversarial classifier with a gradient reversal layer to learn shared representations that remain invariant to the types of questions asked during the interview. This approach aims to reduce biased learning and improve the fairness and generalizability of depression detection in diverse clinical interview scenarios. Classification and regression experiments conducted on three real-world interview-based datasets and one synthetic dataset demonstrate the robustness and generalizability of our model.

1 Introduction

Major depressive disorder, commonly known as depression, is a prevalent mental health condition that can have severe consequences, including emotional distress, social withdrawal, and even suicide. The World Health Organization (WHO)[1] reports that more than 300 million people around the world are affected by depression, which significantly affects individuals, families, and society as a whole. Unfortunately, in many communities, factors such as lack of awareness and financial constraints lead to underdiagnosis and undertreatment of depression and other mental health issues (World Health Organization, 2017).

The clinical interview is the standard method for evaluating depressive symptoms (He et al., 2022). Each interview consists of a sequence of question-and-answer (QA) pairs between the interviewer and the participant. In the past decade, multimodal approaches that incorporate multiple data sources, such as audio, transcribed text, and video collected during interviews, have shown improved performance compared to unimodal methods (Gong and Poellabauer, 2017; Al Hanai et al., 2018; Yang et al., 2024; Zhang et al., 2024).

Recent multimodal studies have shown that the incorporation of the interviewer's question as an additional text modality can improve depression detection (Shen et al., 2022; Milintsevich et al., 2023; Chen et al., 2024; Agarwal et al., 2024). However, the high accuracy reported may stem from models learning interviewer question patterns rather than participant responses (Burdisso et al., 2024). For example, a follow-up question such as "Has your mental health improved?" may reflect previous affirmative responses, providing indirect cues about the participant's condition. Some prior work also uses manually selected QA pairs to distinguish depressed individuals (Yang et al., 2017; Agarwal et al., 2024), often ignoring the broader conversational context. Figure 1 illustrates two excerpts from the same interview. In Figure 1a, the exchange reveals clear depressive cues, whereas Figure 1b is less informative—making the former easier for models to flag as depressed based on surface-level signals.

Therefore, despite the impressive performance metrics presented in previous work, we have significant concerns about the generalizability and robustness of these models on other interview-based speech datasets and real-world clinical situations, where interview questions tend to be more generic

[1] https://www.who.int/news-room/fact-sheets/detail/depression

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12169–12188
November 4-9, 2025

and participants may conceal their true feelings or, in some cases, exaggerate their symptoms (Pretorius et al., 2019; Wilson et al., 2011; Mao et al., 2023; Zhang et al., 2025). It is important that models do not rely on these misleading shortcuts for discrimination.

In this paper, we tackle interviewer-induced bias and shortcut learning in depression detection from clinical interviews using a multimodal architecture that combines interviewer questions and participant responses with gated fusion. We introduce Dialogue-based Contextual Positional Encoding for improved understanding of dialogue turns and a lightweight Dialogue Transformer to capture interview dynamics. Additionally, we apply an adversarial regularization strategy that uses large language model (LLM) annotated interviewer question functions as targets for an adversarial classifier, and the classifier is paired with a gradient reversal layer. As a result, our model is trained to accurately predict depression while simultaneously reducing its dependence on potentially biased interviewer questioning patterns. The code has been released.[2]

2 Related Work

2.1 Unimodal Depression Detection

Unimodal approaches in depression detection typically analyze a single data stream, most commonly focusing on the participant. These include processing participant speech audio using techniques such as self-supervised pre-trained models (Zhang et al., 2021) or traditional acoustic features with LSTMs (Du et al., 2023). Textual analyses have explored participant responses from transcribed interviews, often using methods such as graph convolutional networks for semantic understanding (Burdisso et al., 2023), or leveraging social media text by defining symptom patterns with BERT (Nguyen et al., 2022) or using contrastive learning with capsule networks (Liu et al., 2024).

2.2 Multimodal Depression Detection

Multimodal approaches to depression detection typically integrate two or more data streams from participants, such as speech audio, lexical content from transcripts, and visual cues, to achieve more robust assessments than unimodal methods. Researchers have explored various techniques, including ensembling diverse features with topic modeling (Gong and Poellabauer, 2017), using

[2] https://github.com/coolsoda/e-dep/tree/main

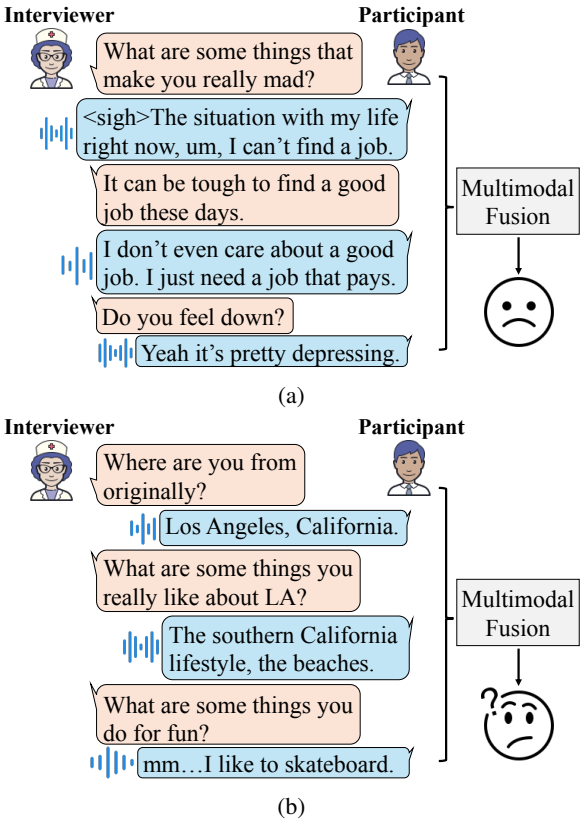

Figure 1: Two segments from the same interview session. The first (top) segment contains more discriminative information for assessing depression, while the second (bottom) is less clear.

LSTMs for sequential modeling of audio-text interactions (Al Hanai et al., 2018), leveraging self-supervised foundational models for enriched representations (Wu et al., 2023), applying expert knowledge in fusion (Yang et al., 2024), using cross-modal attention (Iyortsuun et al., 2024), and integrating acoustic landmarks into LLMs (Zhang et al., 2024).

Recent studies have shown performance gains by including the interviewer's questions as an additional textual modality (Niu et al., 2021; Dai et al., 2021; Shen et al., 2022; Milintsevich et al., 2023; Agarwal et al., 2024; Chen et al., 2024; Xue et al., 2024). However, (Burdisso et al., 2024) found that such models may exploit "shortcuts," relying on interviewer's prompts rather than participant's speech or language, leading to inflated results on certain datasets and poor generalizability across varied interview styles.

Our multimodal framework is designed to counteract the biases introduced by interviewers. We use Dialogue-based Contextual Positional Encoding by merging sequential position with the content

of the questions to get representations that contain turn-level context. Additionally, our adversarial interviewer-behavior regularization, implemented with a gradient reversal layer, trains the model to learn representations that are unaffected by potentially biased questioning patterns. This approach promotes a more robust and genuinely participant-focused assessment of depression.

3 Methodology

3.1 Problem Formulation

The dataset $\mathcal{D}=\{I_1,\ldots,I_n\}$ consists of n clinical interviews, each interview I is composed of a sequence of k question-response pairs $\{(Q_j,A_j)\}_{j=1}^{k}$, where k can vary between interviews. Each question Q_j is presented in text format, while each participant's response A_j is multimodal, including an audio recording A_j^{audio} and its corresponding transcription A_j^{text}. The objective is to predict the depression status for each complete interview I_k. This involves two prediction tasks. The first task is a binary classification, aimed at predicting a depression label $y_k \in \{0,1\}$, where 0 indicates a healthy individual and 1 indicates a depressed individual. The second task is a regression task, in which we predict a continuous depression scalar score $y_k \in \mathbb{R}_{\geq 0}$. The overall framework is illustrated in Figure 2.

3.2 Feature Extraction

Transcribed speech. To extract semantic representations from interviewer questions and participant transcriptions, we use XLM-RoBERTa (**XLMR**) (Conneau et al., 2019), a multilingual transformer-based language model. XLMR is built on the RoBERTa architecture and is trained on 100 different languages using a masked language modeling objective (Conneau et al., 2019). For each input sentence, we first tokenize and encode it to get a fixed-dimensional vector representation. By applying mean pooling over all token embeddings, we derive a sentence-level embedding of a 768-dimensional feature vector for each sentence, denoted q_i^{text} for the question and a_i^{text} for the participant's response.

Audio signal. For the spoken response of the participant, we use Wav2Vec2-XLSR-53 (**XLSR-53**) (Conneau et al., 2020) as a multilingual self-supervised speech encoder to extract audio features. XLSR-53 is a variant of wav2vec 2.0 large (Baevski et al., 2020), pretrained on 53 languages using 16kHz sampled speech audio, including English, simplified Chinese, and Italian. XLSR-53 outputs frame-level audio embeddings with a dimension of 1024. We apply mean pooling over the temporal frames to obtain a fixed-dimensional representation, represented as a_i^{raw}. Then we apply a linear projection layer to map the audio embeddings to the same dimensional space of text (768):

$$a_i^{\text{audio}} = W_{\text{proj}} \cdot a_i^{\text{raw}} + b \tag{1}$$

where $a_i^{\text{raw}} \in \mathbb{R}^{1024}$ is the mean-pooled audio embedding, and $W_{\text{proj}} \in \mathbb{R}^{768\times 1024}$ is the learned projection matrix.

More details on the rationale for choosing XLMR and XLSR-53 over other models, as well as their architecture and pre-training specifics for each language, are discussed in Appendix I.

3.3 Modality Fusion

We adopt a gated fusion mechanism (Arevalo et al., 2017) to combine the interviewer's question (q_j^{text}), the participant's text response (a_j^{text}), and audio features (a_j^{audio}) for each QA pair. Each modality is encoded via a two-layer MLP with ReLU activations:

$$\begin{aligned} h_q &= \text{MLP}_q(q_j^{\text{text}}) \\ h_t &= \text{MLP}_t(a_j^{\text{text}}) \\ h_a &= \text{MLP}_a(a_j^{\text{audio}}) \end{aligned} \tag{2}$$

Each MLP maps $\mathbb{R}^{768} \to \mathbb{R}^{768}$. Gating coefficients are computed using sigmoid activations over linear projections of the original inputs:

$$g_m = \sigma(W_m m_j + b_m), \quad m \in \{q,t,a\} \tag{3}$$

The final fused representation z_j is an element-wise gated sum:

$$z_j = g_q \odot h_q + g_t \odot h_t + g_a \odot h_a \tag{4}$$

3.4 Contextual Positional Encoding

Contextual Positional Encoding (CoPE) (Golovneva et al., 2024) is a method used in transformer-based natural language processing that adjusts position embeddings based on context, rather than fixed absolute or relative indices (Vaswani, 2017; Shaw et al., 2018; Raffel et al., 2020). While effective in tasks such as counting and selective copy, original CoPE, which increments positions based on content tokens, is less suitable for turn-level dialogue modeling.

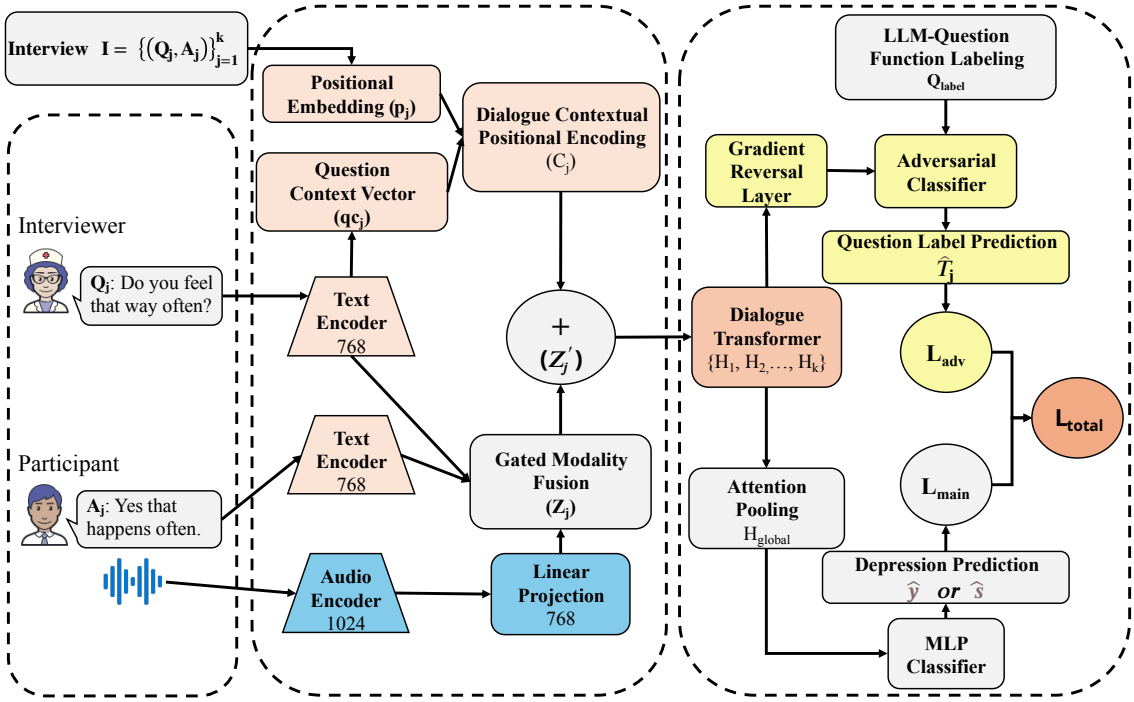

Figure 2: Overview of the proposed multimodal framework for depression detection. (k) is the length of the current interview, and (j) is the index of the turn. The model processes interviewer questions (Q_j) and participant responses (text A_j^t, audio A_j^a) using respective text and audio encoders. Fused representations (Z_j) are added with Dialogue-based Contextual Positional Encoding (C_j) to produce Z'_j. A Dialogue Transformer then generates contextualized turn embeddings H_j. For depression prediction, the H_j sequence is aggregated via attention pooling (H_{global}) and fed to an MLP classifier. In parallel, an adversarial branch with a Gradient Reversal Layer uses H_j to predict LLM-annotated Interviewer Question Functions (Q_{label}/j), encouraging representations invariant to interviewer's question.

To address this, we propose **Dialogue-based CoPE (D-CoPE)**, tailored for encoding question-answer (QA) turns in clinical interviews. D-CoPE integrates both turn position and interviewer semantics. For each QA pair j in an interview of length k, we generate an absolute sinusoidal positional embedding $p_j \in \mathbb{R}^{768}$. The interviewer's question q_j^{text} is passed through a two-layer MLP to extract a contextual vector:

$$qc_j = \text{MLP}_{\text{CoPE}}(q_j^{\text{text}}) \tag{5}$$

We then concatenate p_j and qc_j, and project the result back to the model's hidden space:

$$c_j = W_{\text{CoPE}}[p_j; qc_j] + b_{\text{CoPE}} \tag{6}$$

where $W_{\text{CoPE}} \in \mathbb{R}^{768\times1536}$ and $b_{\text{CoPE}} \in \mathbb{R}^{768}$. The final D-CoPE vector c_j is added element-wise to the fused multimodal representation z_j:

$$z'_j = z_j + c_j \tag{7}$$

3.5 Dialogue-level Transformer

To capture the complete conversational context and model the interdependencies between different QA turns in an interview, we process the sequence of CoPE-enhanced fused representations $\{z'_1, z'_2, \ldots, z'_k\}$ as input for a lightweight two-layer transformer encoder. The sequence $Z' \in \mathbb{R}^{k\times768}$ is fed into a standard transformer encoder with L layers:

$$Z' \in \mathbb{R}^{k\times768} \tag{8}$$

The output of the encoder is a sequence of contextualized representations:

$$H = \text{TransformerEncoder}(Z') \tag{9}$$

where $H \in \mathbb{R}^{k\times768}$. Each vector H_j is a contextualized representation of the j-th QA pair.

3.6 Depression Detection

To obtain a fixed size representation for an entire interview from the sequence of contextualized QA embeddings $\{H_1, \ldots, H_k\}$, we apply attention pooling. The global interview embedding H_{global}

is computed as a weighted sum:

$$H_{\text{global}} = \sum_{j=1}^{T_k} \alpha_j H_j \tag{10}$$

where α_j denotes the learned attention weight for the j-th QA pair. This effectively summarizes all turns within a single interview.

H_{global} is then fed into a two-layer MLP classifier with ReLU activation in the hidden layer. For the binary classification task, the final layer uses a sigmoid activation to produce a probability $\hat{y} \in [0, 1]$ that indicates the likelihood of depression. The model is trained with binary cross-entropy loss:

$$\mathcal{L}_{\text{main}} = -\left[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right] \tag{11}$$

For the regression task, the output layer uses linear activation to produce a scalar prediction $\hat{s}$ corresponding to the severity score of depression. The training objective is the mean squared error (MSE) loss:

$$\mathcal{L}_{\text{main}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{s}_i - s_i)^2 \tag{12}$$

where s_i is the ground truth score for the i-th interview and $\hat{s}_i$ is the model's prediction.

3.7 Prompt Label Prediction

During data preprocessing, for each prompt Q_j from the interviewer, we use LLMs to categorize the prompt into one of seven carefully defined question functions (QFs): 'open-ended', 'change talk', 'neutral information gathering', 'transitional', 'specific probing', 'supportive', and 'other'. These QFs are based on foundational taxonomies for question classification and functional taxonomies from clinical psychology and psychotherapy (Trzepacz and Baker, 1993; Choi and Pak, 2004; Peräkylä et al., 2008; Kallio et al., 2016). Detailed prompts, instructions, and principles followed by the LLM are discussed in Appendix H.

For each QA-level representation H_j, which is the output of the Dialogue-Level Transformer, we train an adversarial classifier. This classifier is preceded by a gradient reversal layer (GRL) and is designed to predict the QF label T_j:

$$\hat{T}_j = \text{MLP}_{\text{adv}}(\text{GRL}(H_j)) \tag{13}$$

MLP_{adv} is a shallow feedforward network with two layers. The adversary tries to predict the QF (T_j) from H_j and minimizes the adversarial loss $\mathcal{L}_{\text{adv}}$, so the gradients of $\mathcal{L}_{\text{adv}}$ flow back to the adversary, making it better at predicting the QF from H_j.

The GRL lies between the main depression prediction model and the adversary. During the backpropagatiom, when gradients from $\mathcal{L}_{\text{adv}}$ reach the GRL, it flips their sign. Therefore, the primary depression prediction model receives two sets of gradient signals to guide the formation of H_j. From the loss function $\mathcal{L}_{\text{main}}$, the model learns to improve H_j to predict depression. In contrast, the loss function $\mathcal{L}_{\text{adv}}$ aims to make H_j less effective in predicting QF. This dual approach requires the main model to identify features for depression prediction that are independent of the QF information.

3.8 Training Objective

The overall training objective for our model is designed to achieve two goals simultaneously: accurately predict the primary outcome (e.g., depression status) and reduce the model's reliance on potentially biased Interviewer Question Functions (QFs). This is accomplished by combining the main prediction loss ($\mathcal{L}_{\text{main}}$) with the adversarial regularization loss ($\mathcal{L}_{\text{adv}}$) with hyperparameter λ:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} - \lambda \cdot \mathcal{L}_{\text{adv}} \tag{14}$$

The hyperparameter λ is a nonnegative scalar that controls trade-off between minimizing the main task's prediction error and minimizing the information about QFs in the learned representations H_j.

4 Experimental Setup

4.1 Data Augmentation and Preprocessing

In this study, we used three real-world datasets and one synthetic dataset, and all datasets are fully anonymized. The Distress Analysis Interview Corpus/Wizard-of-Oz (**DAIC-WOZ**) dataset[3] (Gratch et al., 2014) is one of the most widely used resources for speech-based depression research. This English-language dataset comprises 189 interviews collected from 189 participants. It features 16kHz audio recordings, transcribed text, and manually extracted visual data. Each participant is assigned a score based on the Patient Health Questionnaire with 8 items (PHQ-8) (Kroenke et al., 2009). Of the 189 interviews, 57

[3] https://dcapswoz.ict.usc.edu/

participants are classified as depressed, while 132 are classified as healthy controls. The Emotional Audio-Textual Depression (**EATD**) corpus[4] (Shen et al., 2022) is a Chinese dataset containing interviews with 162 participants, each labeled using the Self-Rating Depression Scale (SDS) (Zung, 1965). In this dataset, 30 participants are identified as depressed, while 132 are classified as healthy controls. It also includes audio recordings at 16 kHz and their corresponding transcriptions. The **Androids** corpus[5] (Tao et al., 2023) consists of 116 interviews with 116 participants, conducted in Italian and labeled according to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (Association et al., 2013). In this dataset, 62 participants are identified as depressed and 54 as healthy controls. The audio recordings are initially at 44.1 kHz and have been resampled to 16 kHz. Detailed information of the three datasets is discussed in Appendix C.

Clinical datasets often exhibit imbalanced class distributions, which can significantly impact model performance. For the DAIC-WOZ dataset, we applied random oversampling in the training set to create duplicate samples for the minority (depressed) class, thus achieving a balance between the number of depressed individuals and healthy controls. For the EATD corpus, we followed the oversampling strategy proposed by the original authors to rearrange the order of QA pairs within each interview, generating additional data points to balance the class distribution (Shen et al., 2022). In both the DAIC-WOZ and EATD datasets, random oversampling was strictly applied to the training set to ensure realistic validation and to prevent any data leakage. The Androids corpus was originally constructed in a balanced manner, so oversampling was not applied.

Given the limited size of the datasets, we use LLM (GPT-4o) (Hurst et al., 2024) to synthesize additional text data for training, following similar approaches as in previous studies (Chen et al., 2024). Synthetic data are based on each training set from the DAIC-WOZ dataset (referred to as **DAIC-Synthetic**) due to its size and common use in this field for research and benchmarking challenges. For each interview in the training set, the interviewer's questions remain unchanged, and the LLM is instructed to generate alternatives of participant responses in text form. These responses are rephrased to maintain the original content and the usage of the vocabulary. We create new samples of participants' audio based on the original audio by random frame-swapping. Since raw video data are not available in the DAIC-WOZ dataset, we duplicated the existing low-level visual features for each interview when implementing the baseline models that utilized the visual modality. Detailed instructions and the prompts provided to the LLM can be found in Appendix B.

[4] https://github.com/speechandlanguageprocessing/ICASSP2022-Depression

[5] https://github.com/androidscorpus/data

4.2 Implementation Details

The proposed framework was implemented using PyTorch. The model was trained and evaluated on Google Colab Pro using an NVIDIA A100 GPU with driver version 550.90.12, CUDA version 12.4, system RAM of 83.5 GB, and GPU RAM of 40 GB. The model was trained with a batch size of 8 for 20 epochs. using the AdamW optimizer (Kingma, 2014) with an initial learning rate of 1×10^{-3}.

Most hyperparameters for each module are discussed in Section 3. A comprehensive list of all hyperparameters and their explored ranges for each module can be found in Appendix F.

4.3 Evaluation

To maintain consistency and allow for a fair comparison with baselines in each dataset, we used five-fold stratified cross-validation for evaluation. For each dataset, all data points are divided into five folds using stratified splitting to ensure that each fold maintains the same class distribution (i.e., the proportion of depressed and healthy participants) as the original dataset. In each training round, one fold is used as the test set, and the remaining four folds are used for training. This process is repeated five times, each fold serving as a test set once. For each training set, we applied the oversampling strategies discussed in Section 4.1 to balance the class distributions. The final performance is reported as the average across the five independent train-test evaluations.

We evaluate the model using four metrics to assess both its classification and regression performance. For the binary classification task of determining whether a participant belongs to the positive class (depressed), we report the F1 score, which balances precision and recall, as well as the AUC-ROC score to evaluate the model's ability to distinguish between the classes across different decision thresholds. For the regression task of predicting

Model	Modality	DAIC-WOZ		EATD		Androids		DAIC-Synthetic	
		F1	RMSE	F1	RMSE	F1	RMSE	F1	RMSE
GCN-PE (Burdisso et al., 2024)	I	**0.84**	4.51	0.55	5.94	0.59	–	**0.84**	4.45
WU (Wu et al., 2023)	A + T	0.80	<u>4.36</u>	0.67	4.21	0.71	–	0.81	<u>4.29</u>
MMPF (Yang et al., 2024)	A + T	0.76	5.11	0.66	4.46	0.70	–	0.78	5.02
ACMA (Iyortsuun et al., 2024)	A + T	0.69	5.15	0.61	5.12	0.68	–	0.72	4.80
LLM (Zhang et al., 2024)	A + T	0.76	5.04	0.64	4.68	0.69	–	0.79	4.61
CAMFM (Xue et al., 2024)	A + T	0.80	4.71	<u>0.70</u>	<u>4.10</u>	<u>0.72</u>	–	0.81	4.51
DAI (Dai et al., 2021)	I + A + T + V	0.81	4.67	0.67	4.33	0.71	–	0.80	4.65
HCAG (Niu et al., 2021)	I + A + T	0.80	4.79	0.65	5.54	0.69	–	0.81	4.43
SHEN (Shen et al., 2022)	I + A + T + V	0.77	5.25	0.68	4.59	0.70	–	0.77	5.14
MILI (Milintsevich et al., 2023)	I + T	0.75	5.11	0.62	5.97	0.68	–	0.76	4.96
SEGA (Chen et al., 2024)	I + A + T + V	0.71	5.04	<u>0.70</u>	4.93	0.70	–	0.74	4.90
GCN (Burdisso et al., 2023)	I + T	0.79	4.95	0.61	5.52	0.69	–	0.80	4.87
AGAR (Agarwal et al., 2024)	I + T	0.72	5.96	0.58	5.53	0.67	–	0.72	5.91
Dialogue Transformer (ours)	I + A + T	<u>0.82</u>	**3.86**	**0.72**	**4.04**	**0.73**	–	<u>0.83</u>	**3.84**

Table 1: Results from training and testing on each dataset individually. *A*, *T*, and *V* refer to participant audio, transcribed text, and visual features; *I* denotes interviewer prompts (text). Best results are in **bold**, second-best are <u>underlined</u>, and – indicates unavailable data. Higher F1 and lower RMSE indicate better performance.

PHQ-8 scores, we use Root Mean Squared Error (RMSE) to measure the magnitude of prediction errors, and Mean Absolute Error (MAE) to offer a complementary measure of average deviations between the predicted and actual scores that accounts for the scale of the values. A detailed calculation for each metric is provided in Appendix D.

4.4 Baseline and Ablations

We compare our method with baselines from three perspectives. First, we include multimodal models that explicitly incorporate interviewer questions, as these questions provide essential context for interpreting participant responses. Second, we consider strong multimodal baselines that exclude interviewer input. Third, we include one study that, to our knowledge, achieves state-of-the-art performance using only interviewer questions on the DAIC-WOZ benchmark. All baselines follow the same training and evaluation setup described in Section 4.3, and each baseline model is discussed in detail in Appendix G.

Our ablation studies are divided into two categories. The first focuses on ablating key modules in our model for the depression detection task, and the second assesses the contributions of each individual data modality.

5 Results and Discussion

5.1 Overall Performance

We compared our method with three baseline groups, as described in Section 4.4, and we present the results in Table 1. Our observations indicate that GCN-PE, a method specifically designed to use cues from interviewer questions, achieves the highest F1 scores in the DAIC-WOZ and DAIC-Synthetic datasets. However, its performance significantly declines on the Androids dataset and even more on the EATD dataset. These two datasets consist of shorter interviews with more generic and less structured prompts from the interviewer, suggesting that GCN-PE lacks generalizability in diverse interview settings.

We found that multimodal approaches that include the interviewer's questions (modality *I*) generally perform well on the DAIC-WOZ and DAIC-Synthetic datasets. However, their performance decreases—sometimes significantly—when applied to EATD and Androids. This indicates that the interviewer's questions in DAIC-WOZ provide more informative and consistent cues than those in the other datasets. Consequently, models that depend on this modality tend to struggle in scenarios where the interviewer's behavior is more generic or varied. This trend is also reflected in our proposed method, where performance noticeably drops on EATD and Androids compared to DAIC-WOZ.

For methods that do not use interviewer prompts,

Model Variant	Modality	DAIC-WOZ		EATD		Androids		DAIC-Synthetic	
		F1	RMSE	F1	RMSE	F1	RMSE	F1	RMSE
D-CoPE + QF + GRL (**Full Model**)	I + A + T	0.82	3.86	**0.72**	4.04	**0.73**	–	0.83	3.84
D-CoPE + QF	I + A + T	**0.89**	**3.81**	0.71	4.12	**0.70**	–	**0.90**	3.82
D-CoPE + GRL	I + A + T	0.83	3.86	**0.72**	**4.01**	**0.73**	–	0.84	**3.74**
QF + GRL	I + A + T	0.65	5.41	0.54	4.90	0.58	–	0.63	5.37
Base Model	I + A + T	0.65	5.57	0.53	4.95	0.57	–	0.67	5.40
Full Model	I + A	0.70	4.52	0.61	5.38	0.63	–	0.74	4.31
Full Model	I + T	0.77	4.34	0.65	4.99	0.66	–	0.79	4.19
Full Model	A + T	0.70	4.50	0.70	4.17	0.65	–	0.72	4.56

Table 2: Ablation study results for depression detection across four datasets. **D-CoPE** refers to our proposed Dialogue-based Contextual Positional Encoding. **QF** indicates the use of Interviewer Question Function labels as targets for the adversarial classifier, and **GRL** refers to the Gradient Reversal Layer used in the adversarial training. 'Base Model' lacks all three of these components but uses the same underlying multimodal architecture. *I*, *A*, and *T* represent the interviewer's prompt in text format , the participant's response in audio, participant's response in text, respectively.

we observe slightly better performance on the DAIC-Synthetic dataset. This improvement is expected due to the more balanced training data and the increased number of samples available. However, for models that use interviewer prompts (*I*) and visual cues (*V*), there is no significant improvement in performance in DAIC-Synthetic. This suggests that the synthesized (i.e., duplicated) visual features provide minimal benefit and may even hinder performance. Therefore, more effective techniques are needed to synthesize low-level visual features.

Regarding the Androids corpus, since it uses DSM-5 for diagnosing depression, it only provides binary labels (depressed or not) (as discussed in Section 4.1). As a result, regression analysis cannot be conducted, as continuous target values are not available. Appendix E contains additional experimental results of the model's performance on the AUC-ROC and MAE metrics. To further assess the performance of the model across varied interviewer styles, domains, and language variations, we performed a comprehensive cross-dataset validation. The results are reported in Table 4 in Appendix E.

5.2 Ablation Studies

The ablation results are summarized in Table 2. D-CoPE plays a critical role: Eliminating it led to sharp declines in both F1 and RMSE, particularly on structured interview datasets such as DAIC-WOZ and Androids. This confirms that D-CoPE effectively encodes both sequential and question-level semantics for the Dialogue-Level Transformer.

Removing the GRL allowed the model to exploit correlations with QFs, resulting in the highest F1 and near-best RMSE scores on DAIC-WOZ and DAIC-Synthetic, surpassing even the full model. This suggests that QFs act as strong shortcut features in these datasets. While removing GRL improves performance on test sets sharing these biases, our full model, by including GRL, intentionally reduces the reliance on such shortcuts. Although this comes at a modest cost in raw performance, it promotes generalization and fairness, which are not fully reflected in the test metrics.

Interestingly, the variant *CoPE + GRL (No QF)*, which omits the adversarial loss against QF labels, performs similarly or slightly better than the full model. This implies that adversarial training introduces useful regularization but may slightly reduce in-distribution accuracy in favor of debiasing.

Modality ablations further demonstrate the contribution of each input stream. Removing either participant text or audio reduced performance across all datasets, with text generally being more critical. Eliminating interviewer questions notably degraded performance on DAIC-WOZ and Androids, highlighting their role in contextualizing participant responses and supporting D-CoPE. On the EATD dataset, which uses generic prompts, removing interviewer input had a less pronounced effect.

6 Conclusion

In this work, we developed a multimodal framework to enhance the detection of depression from clinical speech interview data. The primary goal was to ensure that the model learns meaningful representations of depression from the participants' responses, rather than relying on superficial cues from the interviewers' questions or overfitting to specific, manually-probed question-answer pairs, which can lead to misleading performance improvements. The results highlight the importance of explicitly modeling and addressing potential biases that may arise during the collection of clinical data. We hope that our work contributes to building a more reliable solution for developing fair and generalizable AI systems in mental health assessment.

Limitations

We discuss the limitations of our work from two perspectives: one related to the data used and the other to the model architecture.

- **Data.** There are two main limitations from a data point of view. The first is the size of the available datasets. Labeled clinical speech datasets are often small and constrained by privacy regulations, limiting public access. This challenge extends beyond depression research to other mental and neurodegenerative disorders. To address this, we applied oversampling and used synthetically generated data to improve generalization. However, future work would benefit from larger, more diverse, and publicly available datasets, either collected in real-world settings or generated via advanced models.

  The second limitation is the scope of the modality. Our study focuses on text and audio due to availability constraints. The EATD and Androids datasets lack video data entirely, and while DAIC-WOZ includes some visual features, the original videos are not accessible. Thus, implementing a visual pipeline comparable to our text/audio design was infeasible. Moreover, our work centers on mitigating interviewer bias in audio-textual data, making additional modalities beyond our scope. However, we acknowledge the value of integrating visual (e.g., facial expressions, medical imaging) and physiological signals (e.g., heart rate, skin conductance) to enrich multimodal models for depression detection. This remains a promising direction as more comprehensive datasets emerge.

- **Model.** Our framework is tailored to dyadic clinical interviews, leveraging Dialogue-based Contextual Positional Encoding and adversarial regularization targeting Interviewer Question Functions (IQFs). While effective for structured interviews, the model may not generalize to non-interactive contexts like monologues or self-reports, where key components lose functionality. Adapting our design to such formats presents a separate research challenge.

  IQF definitions are central to our bias mitigation strategy. We grounded these categories in clinical interviewing and dialogue act taxonomies, using LLMs to label questions in context. While this approach enables scalable annotation, it may not fully capture the nuance of real interviewer behavior. LLM outputs also depend on the underlying model and prompting strategy, which may limit precision. Future work can use human annotators to cover a broader range of interviewer behavior labels and help verify the accuracy of the labels produced by LLM, using available resources.

References

Navneet Agarwal, Gaël Dias, and Sonia Dollfus. 2024. Analysing relevance of discourse structure for improved mental health estimation. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 127–132.

Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*. American psychiatric association.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Sergio Burdisso, Ernesto Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, Pastor López-Monroy, and Petr Motlicek. 2024. Daic-woz: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews. *arXiv preprint arXiv:2404.14463*.

Sergio Burdisso, Esaú Villatoro-Tello, Srikanth Madikeri, and Petr Motlicek. 2023. Node-weighted graph convolutional network for depression detection in transcribed clinical interviews. *arXiv preprint arXiv:2307.00920*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Zhuang Chen, Jiawen Deng, Jinfeng Zhou, Jincenzi Wu, Tieyun Qian, and Minlie Huang. 2024. Depression detection in clinical interviews with llm-empowered structural element graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8174–8187.

Bernard CK Choi and Anita WP Pak. 2004. A catalog of biases in questionnaires. *Preventing chronic disease*, 2(1):A13.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

Zhijun Dai, Heng Zhou, Qingfang Ba, Yang Zhou, Lifeng Wang, and Guochen Li. 2021. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. *Journal of affective disorders*, 295:1040–1048.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Minghao Du, Shuang Liu, Tao Wang, Wenquan Zhang, Yufeng Ke, Long Chen, and Dong Ming. 2023. Depression recognition using a proposed speech chain model fusing speech production and perception features. *Journal of Affective Disorders*, 323:299–308.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. 2024. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*.

Yuan Gong and Christian Poellabauer. 2017. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, pages 69–76.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Reykjavik.

Lang He, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, Hongyu Wang, Songtao Ding, Zhongmin Wang, et al. 2022. Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, 80:56–86.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ngumimi Karen Iyortsuun, Soo-Hyung Kim, Hyung-Jeong Yang, Seung-Won Kim, and Min Jhon. 2024. Additive cross-modal attention network (acma) for depression detection based on audio and textual features. *IEEE Access*.

Hanna Kallio, Anna-Maija Pietilä, Martin Johnson, and Mari Kangasniemi. 2016. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of advanced nursing*, 72(12):2954–2965.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.

Han Liu, Changya Li, Xiaotong Zhang, Feng Zhang, Wei Wang, Fenglong Ma, Hongyang Chen, Hong Yu, and Xianchao Zhang. 2024. Depression detection via capsule networks with contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22231–22239.

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*.

Kaining Mao, Yuqi Wu, and Jie Chen. 2023. A systematic review on automated clinical depression diagnosis. *Npj mental health research*, 2(1):20.

Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2023. Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, 10(1):4.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. *arXiv preprint arXiv:2204.10432*.

Meng Niu, Kai Chen, Qingcai Chen, and Lufeng Yang. 2021. Hcag: A hierarchical context-aware graph attention model for depression detection. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4235–4239. IEEE.

Anssi Peräkylä, Charles Antaki, Sanna Vehviläinen, and Ivan Leudar. 2008. Analysing psychotherapy in practice. In *Conversation analysis and psychotherapy*, pages 5–25. Cambridge University Press.

Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur. 2014. Parallel training of dnns with natural gradient and parameter averaging. *arXiv preprint arXiv:1410.7455*.

Claudette Pretorius, Derek Chambers, and David Coyle. 2019. Young people's online help-seeking and mental health difficulties: Systematic narrative review. *Journal of medical Internet research*, 21(11):e13873.

James O Prochaska and Wayne F Velicer. 1997. The transtheoretical model of health behavior change. *American journal of health promotion*, 12(1):38–48.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251. IEEE.

John Sommers-Flanagan and Rita Sommers-Flanagan. 2012. *Clinical interviewing: 2012-2013 update*. John Wiley & Sons.

Fuxiang Tao, Anna Esposito, and Alessandro Vinciarelli. 2023. The androids corpus: A new publicly available benchmark for speech based depression detection. *Depression*, 47:11–9.

Paula T Trzepacz and Robert W Baker. 1993. *The psychiatric mental status examination*. Oxford University Press.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. *arXiv preprint arXiv:2012.04080*.

Coralie J Wilson, Debra J Rickwood, John A Bushnell, Peter Caputi, and Susan J Thomas. 2011. The effects of need for autonomy and preference for seeking help from informal sources on emerging adults' intentions to access mental health services for common mental disorders and suicidal thoughts. *Advances in Mental Health*, 10(1):29–38.

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

Wen Wu, Chao Zhang, and Philip C Woodland. 2023. Self-supervised representations in speech-based depression detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Junqi Xue, Ruihan Qin, Xinxu Zhou, Honghai Liu, Min Zhang, and Zhiguo Zhang. 2024. Fusing multi-level features from audio and contextual sentence embedding from text for interview-based depression detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6790–6794. IEEE.

Biao Yang, Miaomiao Cao, Xianlin Zhu, Suhong Wang, Changchun Yang, Rongrong Ni, and Xiaofeng Liu. 2024. Mmpf: Multimodal purification fusion for automatic depression detection. *IEEE Transactions on Computational Social Systems*.

Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2017. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, pages 53–59.

Enshi Zhang, Rafael M Trujillo, and Christian Poellabauer. 2025. Participant engagement and data quality: Lessons learned from a mental wellness crowdsensing study. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–30.

Pingyue Zhang, Mengyue Wu, Heinrich Dinkel, and Kai Yu. 2021. Depa: Self-supervised audio embedding for depression detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 135–143.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. 2024. When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. *arXiv preprint arXiv:2402.13276*.

William WK Zung. 1965. A self-rating depression scale. *Archives of general psychiatry*, 12(1):63–70.

Appendix

A Overview

The appendix is organized as follows. Section B provides details on how we used LLM to synthesize additional data (DAIC-Synthetic) for training and evaluation purposes. Section C presents more information about each real-world dataset used in this work. Section D lists the main evaluation metrics employed in this study. Section E includes experiments conducted on two additional metrics, as well as cross-dataset validation across the four datasets. Section F outlines the detailed hyperparameters for general training and each module within the framework. Section G describes the implementation details of the baseline models. Section H discusses how we used the LLM to annotate the question functions, and Section I presents the rationale behind the foundational models used for feature extraction in this work.

B Synthetic Data

In Figure 3, we present the prompt template used to instruct the LLM through GPT-4 API calls to synthesize participant responses based on interviews from the DAIC-WOZ dataset. Each interview consists of multiple QA pairs. For each QA pair, we prompt the LLM to generate three alternative text responses.

For model training and evaluation (as discussed in Section 4.1), we use a 5-fold cross-validation strategy with stratified splitting to ensure that each fold retains the original class distribution. During each fold, the training set, which comprises 80% of the data, is augmented by synthesizing additional samples. This augmentation not only increases the size of the training set but also improves class balance, ultimately enhancing the model's ability to generalize. After augmentation, the size of each training set increases from 152 to 396 samples.

C Details of Data Collection and Preprocessing

The Distress Analysis Interview Corpus/Wizard-of-Oz (**DAIC-WOZ**) dataset (Gratch et al., 2014) is one of the most widely used datasets for depression studies. This English-language dataset includes 189 interviews conducted with 189 participants and collects 16 kHz audio recordings, transcribed text, and visual data. It was collected from two groups of participants in the Greater Los Angeles area. One group consists of veterans of the U.S. armed forces, while the other includes individuals from the general public. Each interview involves a participant and a human-controlled agent named Ellie. Each participant is labeled with a Patient Health Questionnaire of 8 items (PHQ-8) (Kroenke et al., 2009) score, which ranges from 0 to 24. Participants who score 10 or more are classified as depressed, while those with scores below 10 are classified as healthy controls. Of the 189 interviews, 57 participants are labeled depressed, and 132 are classified as healthy controls.

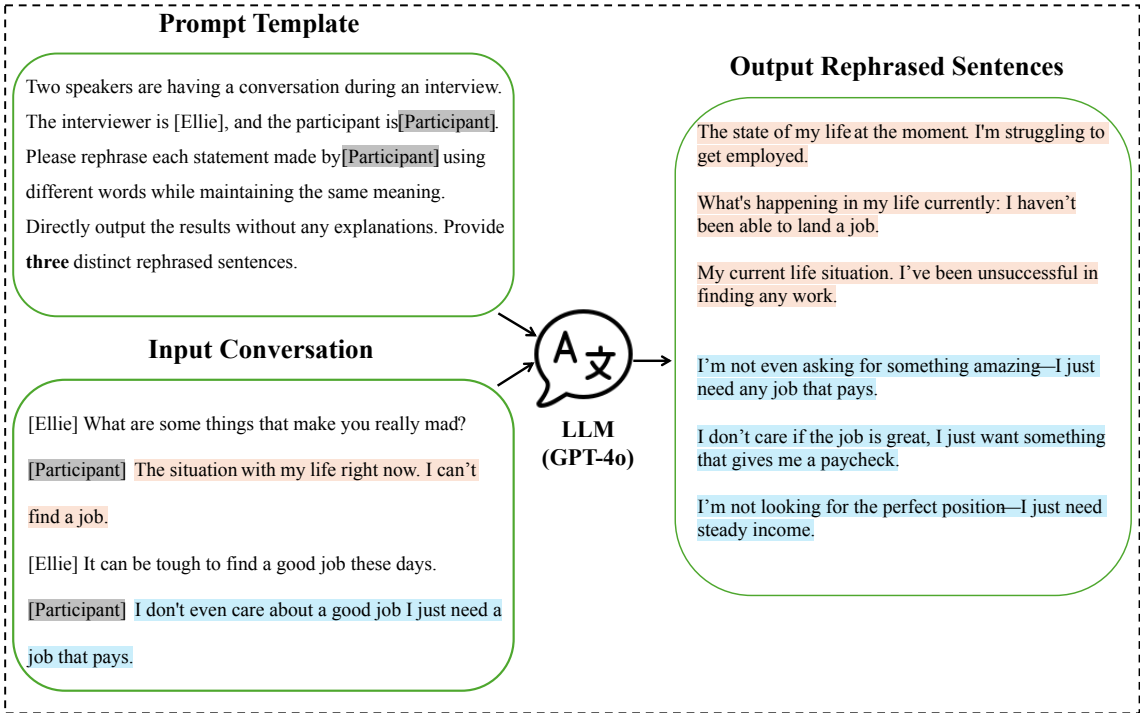

Figure 3: Use LLM to generate synthetic data by rephrasing responses from participants for each question-answer pair.

The Emotional Audio-Textual Depression (**EATD**) corpus (Shen et al., 2022) is a Chinese dataset consisting of interviews with 162 student volunteers. One motivation for its development is to detect depression by studying responses from participants to random, less specifically designed mental health assessment questions. The interviews were recorded as audio at a frequency of 16 kHz and transcribed into text using Kaldi (Povey et al., 2014), which are then manually checked and corrected. The interview questions were not recorded in audio format and were developed based on the participants' responses. Participants were asked to respond to three randomly selected questions from a large pool of questions and to complete a Self-Rating Depression Scale (SDS) questionnaire (Zung, 1965). This questionnaire consists of 20 items, with total raw scores ranging from 20 to 80. In clinical practice, these scores are often converted into an SDS index (raw score multiplied by 1.25) for standardized interpretation, resulting in a range of 25 to 100. For binary classification purposes, a cutoff score of 63 is used; scores of 63 or higher indicate depression. Among the 162 participants, 132 were classified as non-depressed, while 30 were classified as depressed.

The **Androids** Corpus was collected through the ANDROIDS project (Tao et al., 2023) to study depression from participants' speech. It includes 118 native Italian speakers, and 116 of them have undergone clinical interviews. There are 116 audio files available, and we resampled them from 44.1 to 16 kHz. We used Whisper (Radford et al., 2023) to transcribe the audio recordings, capturing both the interviewer's questions and the participants' responses for each interview. Among these 116 individuals, 64 were diagnosed with depression, while 52 were healthy controls. Participants were labeled as depressed or not by medical professionals based on the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (Association et al., 2013). A detailed description of the reported symptoms is not available.

Each interview consists of a varying number of Question-Answer (QA) pairs, and the duration of each participant's spoken response (A_j^q) within these pairs also differs. Our audio processing pipeline first segments each interview into individual QA turns. For each participant's spoken response within a QA turn, which is sampled at 16 kHz, the audio is fed into the Wav2Vec2-XLSR-53 audio encoder. This model features a convolutional layer that processes the audio, outputting a sequence of local acoustic feature vectors every 20 milliseconds. These features are then processed by the model's Transformer layers, resulting in a se-

quence of frame-level embeddings, each with 1024 dimensions. The number of frames in this sequence corresponds directly to the duration of the specific audio utterance.

To create a fixed-size representation for each variable-length audio utterance, we apply temporal mean pooling across all frame-level embeddings for A_j^a. This process aggregates the entire sequence of 1024-dimensional embeddings into a single fixed-dimensional vector. This vector is then linearly projected to 768 dimensions to ensure alignment with our text feature dimensionality before performing multimodal fusion.

After modality fusion and D-CoPE enhancement, we obtain a sequence of fixed-size representations for all question-answer (QA) turns in an interview, denoted as $\{z'_1, z'_2, \ldots, z'_k\}$. To manage the varying number of QA pairs (k) across different interviews when batching for the Dialogue-Level Transformer, these sequences are zero-padded to a uniform maximum sequence length. This maximum length is determined based on the 95th percentile of interview lengths, measured by the number of QA pairs. Specifically, this maximum length is set to 35 for the DAIC and DAIC-Synthetic datasets, 3 for the EATD dataset, and 26 for the Androdis Corpus.

D Evaluation Metrics

For binary classification tasks, we primarily use the F1 score. True positives (**TP**) refer to participants correctly predicted as depressed. False positives (**FP**) refer to participants incorrectly predicted as depressed. False negatives (**FN**) refer to participants who are actually depressed but incorrectly classified as non-depressed by the model.

The metrics are defined as follows:

- **Precision** measures the proportion of true positive predictions among all positive predictions made by the model. It is defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

- **Recall** (also known as sensitivity) measures the proportion of true positive predictions among all actual positive instances. It is defined as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

- **F1 Score** is the harmonic mean of precision and recall, providing a balanced measure of both. It is defined as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

- **AUC-ROC** (Area Under the Receiver Operating Characteristic Curve) evaluates the model's ability to discriminate between the positive and negative classes across all possible classification thresholds. A higher AUC indicates stronger overall separability.

For regression tasks, we report error-based metrics that capture the magnitude of prediction deviations from the ground truth:

- **Root Mean Squared Error (RMSE)** is a widely used metric that measures the average magnitude of prediction errors. It is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{18}$$

  where y_i represents the ground truth value, $\hat{y}_i$ is the predicted value, and N is the total number of samples in the test set. A lower RMSE indicates higher predictive accuracy.

- **Mean Absolute Error (MAE)** measures the average absolute difference between predicted and actual values, offering an interpretable metric in the same units as the target variable. It is defined as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{19}$$

  Compared to RMSE, MAE is less sensitive to large errors and provides a complementary view of model performance.

E Additional Experimental Results

Table 3 shows the experimental results for each dataset, focusing on AUC-ROC and MAE.

The cross-dataset validation results presented in Table 4 indicate several key observations. First, the model trained on the DAIC-WOZ dataset transfers reasonably well to the Androids dataset and performs acceptably on the EATD dataset. However, when the model is trained on EATD, its performance is consistently poor across all other datasets.

Model	Modality	DAIC-WOZ		EATD		Androids		DAIC-Synthetic	
		AUC-ROC	MAE	AUC-ROC	MAE	AUC-ROC	MAE	AUC-ROC	MAE
GCN-PE (Burdisso et al., 2024)	I	**0.89**	4.06	0.60	5.35	0.64	–	0.88	4.00
WU (Wu et al., 2023)	A + T	0.84	3.90	0.71	3.79	0.77	–	0.85	3.86
MMPF (Yang et al., 2024)	A + T	0.81	4.60	0.73	4.01	0.76	–	0.82	4.52
ACMA (Iyortsuun et al., 2024)	A + T	0.76	4.64	0.69	4.61	0.72	–	0.74	4.32
LLM (Zhang et al., 2024)	A + T	0.80	4.54	0.70	4.21	0.74	–	0.83	4.15
CAMFM (Xue et al., 2024)	A + T	0.84	4.24	0.77	3.69	0.76	–	0.85	4.06
DAI (Dai et al., 2021)	I + A + T + V	0.84	4.20	0.72	3.90	0.77	–	0.83	4.18
HCAG (Niu et al., 2021)	I + A + T	0.85	4.31	0.70	4.98	0.75	–	0.79	3.99
SHEN (Shen et al., 2022)	I + A + T + V	0.80	4.72	**0.77**	4.13	0.72	–	0.80	4.63
MILI (Milintsevich et al., 2023)	I + T	0.82	4.60	0.65	5.37	0.70	–	0.79	4.46
SEGA (Chen et al., 2024)	I + A + T + V	0.80	4.54	0.77	4.44	0.74	–	0.79	4.41
GCN (Burdisso et al., 2023)	I + T	0.84	4.46	0.69	4.97	0.72	–	0.82	4.38
AGAR (Agarwal et al., 2024)	I + T	0.80	5.36	0.66	4.98	0.75	–	0.80	5.32
Dialogue Transformer (ours)	I + A + T	**0.89**	**3.47**	0.75	**3.64**	**0.80**	–	**0.90**	**3.46**

Table 3: Best results are in **bold**, and – indicates unavailable data. Higher AUC-ROC and lower MAE indicate better performance.

Train \ Test	DAIC-WOZ				EATD				Androids				DAIC-Synthetic			
	F1	RMSE	AUC-ROC	MAE	F1	RMSE	AUC-ROC	MAE	F1	RMSE	AUC-ROC	MAE	F1	RMSE	AUC-ROC	MAE
DAIC-WOZ	**0.82**	**3.86**	**0.89**	**3.47**	0.67	4.47	0.74	4.10	0.72	4.20	0.79	3.87	–	–	–	–
EATD	0.61	5.22	0.69	4.78	**0.72**	**4.04**	**0.75**	**3.64**	0.59	5.10	0.65	4.59	0.61	5.25	0.68	4.97
Androids	0.70	–	0.77	–	0.66	–	0.72	–	**0.73**	–	**0.80**	–	0.69	–	0.76	–
DAIC-Synthetic	–	–	–	–	0.68	4.40	0.74	3.96	0.74	4.10	0.78	3.66	**0.83**	**3.84**	**0.90**	**3.46**

Table 4: Cross-dataset validation results.

Additionally, the model trained on Androids generalizes effectively to both DAIC-WOZ and DAIC-Synthetic, although there is a slight performance decline when tested on EATD. Conversely, the model trained on DAIC-Synthetic performs reasonably well on EATD and achieves even better results on Androids.

These findings suggest that the Androids dataset shares a more similar semi-structured interview format with DAIC-WOZ, while the limited question diversity and shorter session lengths of EATD hinder the model's ability to learn generalizable features from other datasets.

Since the synthetic data was primarily used to augment the dataset for training purposes, we did not perform cross-dataset validation between DAIC-WOZ and DAIC-Synthetic to avoid the risk of overfitting and biases.

F Details of Hyperparameters

In Table 5, we present the hyperparameters for general training and all other modules in our framework.

G Baseline Implementation

In this section, we provide implementation details for all baseline models, as listed in Table 1.

- GCN-PE (Burdisso et al., 2024): This study builds LongBERT (Beltagy et al., 2020) and Graph Convolutional Network (GCN)-based models to analyze participant responses and interviewer prompts. They demonstrated that models that use interviewer prompts can achieve high accuracy, even reaching state-of-the-art performance. A qualitative analysis revealed that these models tend to focus on specific, localized segments of the interview, particularly when interviewer asks targeted questions about the participant's mental health history. This suggests that models learn to use these prompts as shortcuts for differentiation, rather than truly understanding the patient's depressive state through their own language.
- WU (Wu et al., 2023): In this study, foundation models were pre-trained using self-supervised learning (SSL) to address data sparsity in speech-based depression detection (SDD). The primary method involves analyzing SSL representations from various layers of foundation models such as wav2vec 2.0, HuBERT, and WavLM to identify indicators of depression. Subsequently, these models are fine-tuned on tasks related to Automatic Speech Recognition (ASR) and Automatic

Component	Hyperparameter	Values Explored
General Training	optimizer	Adam, **AdamW**
	learning rate	$\{1 \times 10^{-3}, 5 \times 10^{-4}, \mathbf{1 \times 10^{-4}}\}$
	batch size	{4, **8**, 16}
	epochs	{10, **20**, 30}
	early stopping patience	{**5**, 10}
	weight decay	{0, **0.01**, 0.05, 0.1}
Gated Fusion (MLP)	number of layers	{1, **2**}
	hidden units	{512, **768**}
D-CoPE	number of layers	{1, **2**}
	hidden units	{**256**, 512}
Dialogue-Transformer	number of layers	{1, **2**, 3}
	number of attention heads	{4, **8**, 12}
	feedforward network dimension	{1536, **2048**}
	dropout rate	{0.1, **0.15**, 0.2, 0.25}
Adversarial Classifier (MLP)	number of layers	{1, **2**, 3}
	hidden units	{64, **128**, 256}
	lambda λ	{0.01, 0.05, 0.1, **0.25**, 0.5, 0.75, 1, 5}

Table 5: The list of hyperparameters explored for the proposed multimodal depression detection framework is presented. The optimal set, highlighted in **bold**, gives the best average performance across the validation folds.

Emotion Recognition (AER) to facilitate the transfer of knowledge to SDD.

- MMPF (Yang et al., 2024): Proposed a multimodal fusion framework for analyzing depression using audio, video, and text streams from clinical interviews. The system extracts features from each modality, including innovative text descriptors from Paragraph Vector for selected responses and a video descriptor from facial landmarks. These features are processed through a Deep Convolutional Neural Network (DCNN) to learn high-level representations. The learned features are then inputted into a Deep Neural Network (DNN) to predict initial PHQ-8 depression scores for each modality. Finally, the scores from the individual pipelines are combined in a fusion DNN to produce the final multimodal PHQ-8 score prediction.

- ACMA (Iyortsuun et al., 2024): This work processes audio data into Mel spectrograms and text data, represented by participant responses encoded using the Universal Sentence Encoder. Two separate Bidirectional Long Short-Term Memory (BiLSTM) networks are used for this, each followed by an attention layer that captures important unimodal features. These processed unimodal representations are then input into the ACMA network, which utilizes an additive attention mechanism to weigh and combine the cross-modal interactions, learning the relationships between speech and text cues.

- LLM-acoustic (Zhang et al., 2024): The work focuses on three main steps: First, extract discrete acoustic landmarks from speech signals. Second, fine-tuning the model through cross-modal instruction using Low-Rank Adaptation (LoRA). This step teaches the LLM to understand these landmarks and their relationship with text, while incorporating "hints" regarding the speaker's depression status. Finally, we employ P-tuning to train the LLM to integrate both the text and the learned landmark representations for the final task of depression detection.

- CAMFM (Xue et al., 2024): This work presents a multi-modal model for detecting depression that integrates audio features at multiple levels with textual sentence embeddings.

For the audio component, we extracted Low-Level Descriptors (LLDs), mel-spectrograms, and features from the wav2vec model. These features are then combined using a Multi-level Audio Features Interaction Module (MAFIM) to form a comprehensive audio representation. In the text domain, we utilize pre-trained BERT to obtain sentence-level embeddings. These distinct audio and text representations are then fused together using a novel Channel Attention-based Multi-modal Fusion Module, allowing for the integration of heterogeneous data.

- DAI (Dai et al., 2021): This work begins by constructing a high-dimensional feature vector obtained from audio, video, and semantic data through context-aware analysis based on the topics extracted. The core of the proposed method involves a two-stage feature selection algorithm. First, a filter method ranks high-dimensional features to identify an informative subset of candidates. Next, a wrapper method refines this subset by sequentially adding features and utilizing a Support Vector Machine (SVM) model to retain only those features that enhance prediction accuracy.

- HCAG (Niu et al., 2021): This paper presents a Hierarchical Context-Aware Graph Attention Model for depression detection. The model begins by utilizing a Sequential Encoder with Gated Recurrent Units (GRUs) and an additive attention mechanism to generate representations for each question-answer pair from text (using GloVe embeddings) or audio (utilizing MFCCs and eGeMAPS features). Following this, a Subject-Level Context Encoder constructs a graph in which the question-answer pairs act as nodes. A Graph Attention Network (GAT) is then used to aggregate contextual information and learn the relationships among these question-answer pairs within a defined context window.

- SHEN (Shen et al., 2022): Mel spectrograms are extracted from audio recordings and converted into fixed-length audio embeddings using NetVLAD (Arandjelovic et al., 2016). These embeddings are then input into a Gated Recurrent Unit (GRU) network. Meanwhile, sentence embeddings generated from interview transcripts using ELMo are processed by a Bidirectional LSTM (BiLSTM) network that is equipped with an attention mechanism to capture important linguistic cues. The feature representations of the text (BiLSTM) and audio (GRU) branches are concatenated, and a "modal attention" mechanism is applied to weigh their respective contributions. The resulting fused representation is then passed to a fully connected network for the final binary classification of depression.

- MILI (Milintsevich et al., 2023): This work proposes a model that uses a hierarchical architecture to process textual transcripts for the detection of depression. First, a Sentence-RoBERTa model encodes individual dialogue turns. Then, a Bidirectional LSTM (BiLSTM) with an additive attention mechanism processes these turn embeddings to create a comprehensive representation of the interview for prediction.

- SEGA (Chen et al., 2024): This work focuses on using expert knowledge in the assessment of depression by constructing a structural element graph. It establishes a directed acyclic graph in which information flows from auxiliary nodes (audio, video, and questions) to a central node (the answer transcript) within each interview round. Central transcript nodes are linked to capture temporal dependencies, and all central nodes connect to a summary node that represents the semantics of the entire interview. Ultimately, a graph attention network is built to learn from the constructed graph.

- GCN (Burdisso et al., 2023): This work uses Graph Convolutional Networks (GCNs) to classify interview transcripts. The modified GCN features a new weighting scheme for edges, particularly for self-connections, where weights are determined by the PageRank algorithm to reflect the importance of each node (word or document). A heterogeneous graph is created from word nodes (one-hot vectors) and document nodes (TF-IDF features). Connections are based on Point-wise Mutual Information for word-word links and TF-IDF for word-document links. This approach enables modeling of long-distance semantics and effectively classifies subjects as either depressed or in the control group.

Question Function	Definition	Example	Source
'open-ended'	Questions that encourage participants to express themselves freely and broadly about a topic, including their experiences, thoughts, or feelings, typically inviting more than just short or specific answers.	'What activities do you enjoy for fun?'	MI (Miller and Rollnick, 2012), CI (Sommers-Flanagan and Sommers-Flanagan, 2012)
'change talk'	Questions directed at helping the participant express their motivations, reasons, desires, abilities, or needs related to making behavioral, cognitive, or situational changes.	'What got you to seek help?'	MI (Miller and Rollnick, 2012), TMC (Prochaska and Velicer, 1997)
'neutral information gathering'	Questions seeking specific factual details, clarification of objective information, or direct answers to concrete inquiries, often expecting a constrained or brief response.	'How easy is it for you to get a good night's sleep?'	DAMSL (Core and Allen, 1997), CI (Sommers-Flanagan and Sommers-Flanagan, 2012)
'transitional'	Utterances (questions or statements) used by the interviewer to organize the conversation, manage transitions between topics, introduce or close segments, summarize points, or manage interview logistics.	'Do you feel that way often?'	MI (Miller and Rollnick, 2012)
'specific probing'	Follow-up questions designed to elicit more detailed information, elaboration, or specific examples regarding a topic or statement previously introduced by the participant, generally in a neutral, non-leading manner.	'Have you being diagnosed with depression?'	CI (Sommers-Flanagan and Sommers-Flanagan, 2012)
'supportive'	Statements or questions that primarily convey empathy, understanding, validation of the participant's feelings or experiences, offer encouragement, build rapport, or affirm their strengths.	'That sounds incredibly difficult.'	MI (Miller and Rollnick, 2012), ERT (Welivita and Pu, 2020)
'other'	Utterances that do not clearly fit into any of the other defined functional categories. This can include very short backchannels, incomplete or interrupted sentences, unintelligible speech, or off-topic remarks not related to structuring.	'Yeah.'; 'hmm.'	DAMSL (Core and Allen, 1997)

Table 6: A complete list of all the Interview Question Functions (IQF) we defined based on established theories. MI stands for Motivational Interviewing, CI stands for Clinical Interviewing Techniques, TMC represents the Transtheoretical Model of Change, DAMSL refers to Dialog Act Markup in Several Layers, and ERT stands for Empathetic Response Taxonomies.

- AGAR (Agarwal et al., 2024): This work proposes a multi-view architecture aimed at improving automated depression estimation from transcripts of patient-therapist interviews by explicitly considering the structure of the discourse. The central method involves dividing the transcript into two distinct "view"—one for therapist questions and the other for patient answers. Each view is processed by a dedicated View Encoder that utilizes multi-head attention to learn specific representations from sentence-level encodings generated by sentence transformers. Importantly, these View Encoders interact through a cross-attention mechanism, enabling them to learn in a co-dependent manner by sharing attention scores.

H Prompt Labeling

In Table 6, we present a list of all the question functions (QFs) we defined based on established methods and theories in communication research and clinical psychotherapy.

Due to limited resources, we employed three LLMs for annotation: GPT-3.5 Turbo, GPT-4o, and LLaMA3-70B. Along with the prompt template summarized in Table 7, each LLM was provided with the definition and example of each QF, along with the relevant context. The 'context' for the current sentence that needs annotation consists of the preceding conversation, with a context length carefully defined to accommodate our resource limitations; in our study, the context length is set to 5. Therefore, for each interviewer's utterance from each QA pair, the LLM considers information from

Prompt Template
Two speakers, the interviewer and the participant, are engaged in a clinical interview. The conversation is {context}. Now interviewer {interviewer} says: {current sentence}. Predict the question function of the sentence {current sentence} from the options [open-ended, change talk, neutral information gathering, transitional, specific probing, supportive, other], consider the conversation context, do not explain, only output the label in [open-ended, change talk, neutral information gathering, transitional, specific probing, supportive, other].

Table 7: The prompt used to annotate the question function of each utterance expressed by the interviewer.

a maximum of 5 preceding utterances to annotate the current utterance.

To determine the final label, we used a majority vote from the outputs of the three models. If there was no consensus among the models, we labeled the utterance as "Other."

I Foundation Models

The success of our multimodal depression detection framework is heavily reliant on the quality of the initial feature representations obtained from both textual and audio modalities. This section outlines the reasoning behind our selection of encoders.

I.1 Textual Feature Extraction

For processing interviewer questions (Q_j^{text}) and participant transcribed responses (A_j^{text}), we use XLM-RoBERTa (XLMR) (Conneau et al., 2019; Liu et al., 2019) as our text encoder. The key reason is its strong multilingual capability. Our datasets span multiple languages, including English (DAIC-WOZ, DAIC-Synthetic), Mandarin Chinese (EATD), and Italian (Androids). XLMR is pre-trained on 100 languages using a massive multilingual corpus and builds upon the robust RoBERTa architecture, enabling it to produce high-quality, cross-lingually consistent embeddings. This choice allows us to maintain methodological consistency across datasets without resorting to separate language-specific models.

We considered several alternatives:

- **Monolingual models** (for example, BERT-base (Devlin et al., 2018) for English) would require different encoders per language, complicating the system design and potentially introducing cross-lingual inconsistencies.
- **Older multilingual models** such as mBERT (Devlin et al., 2018) are generally outperformed by XLM-R due to its superior pre-training strategy and corpus scale.
- **Large generative language models** (for example, GPT-series) demonstrate excellent language understanding (Hurst et al., 2024), but using them as feature encoders—especially with fine-tuning—would be computationally expensive and impractical for processing numerous short textual segments in our sequential pipeline.

Therefore, XLMR offers a favorable trade-off between multilingual representational power and computational efficiency, making it well suited for sentence-level embedding extraction in our downstream modules.

I.2 Audio Feature Extraction

To encode participant spoken responses (A_j^{a}), we use Wav2Vec2-XLSR-53 (XLSR-53) (Conneau et al., 2020; Baevski et al., 2020). This choice is based on several factors:

- **Multilingual capability.** XLSR-53 is pre-trained on speech from 53 languages, including English, Mandarin Chinese, and Italian. It uses a self-supervised learning approach on raw audio waveforms, making it particularly effective at generating consistent and comparable audio representations across diverse linguistic datasets without language-specific fine-tuning.
- **Learning from raw audio.** Unlike traditional acoustic methods that rely on hand-crafted features such as MFCCs or eGeMAPS (Eyben et al., 2010, 2015), models based on Wav2Vec2 learn representations directly from the raw waveform. This approach aligns with our goal of developing a data-driven, end-to-end trainable system wherever possible.
- **Performance.** Wav2Vec2 and its multilingual variant, XLSR-53, have shown outstanding results across a wide range of speech processing benchmarks, emphasizing the quality and

generalizability of the learned audio representations. Other self-supervised speech models, such as HuBER (Hsu et al., 2021) and WavLM (Chen et al., 2022), are strong candidates as well. However, XLSR-53's explicit focus on cross-lingual capabilities and its empirical success in multilingual speech tasks make it particularly suitable for our multilingual setup.