



HiQuE: Hierarchical Question Embedding Network for Multimodal Depression Detection

Juho Jung
Dept. of Applied Artificial Intelligence
Sungkyunkwan University
Seoul, South Korea
jhjeon9@g.skku.edu

Chaewon Kang
Dept. of Applied Artificial Intelligence
Sungkyunkwan University
Seoul, South Korea
codnjs3@g.skku.edu

Jeewoo Yoon
Raondata
Seoul, South Korea
jeewooyoon@raondata.ai

Seungbae Kim
Computer Science and Engineering
University of South Florida
Tampa, Florida, USA
seungbae@usf.edu

Jinyoung Han*
Dept. of Applied Artificial Intelligence
Sungkyunkwan University
Seoul, South Korea
jinyounghan@skku.edu

Abstract

The utilization of automated depression detection significantly enhances early intervention for individuals experiencing depression. Despite numerous proposals on automated depression detection using recorded clinical interview videos, limited attention has been paid to considering the hierarchical structure of the interview questions. In clinical interviews for diagnosing depression, clinicians use a structured questionnaire that includes routine baseline questions and follow-up questions to assess the interviewee's condition. This paper introduces **HiQuE** (**H**ierarchical **Q**uestion **E**MBEDding network), a novel depression detection framework that leverages the hierarchical relationship between primary and follow-up questions in clinical interviews. HiQuE can effectively capture the importance of each question in diagnosing depression by learning mutual information across multiple modalities. We conduct extensive experiments on the widely-used clinical interview data, DAIC-WOZ, where our model outperforms other state-of-the-art multimodal depression detection models and emotion recognition models, showcasing its clinical utility in depression detection.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Multimedia information systems**; **Data mining**.

Keywords

Multimodal Depression Detection, Hierarchical Question Embedding, Clinical Interview

ACM Reference Format:

Juho Jung, Chaewon Kang, Jeewoo Yoon, Seungbae Kim, and Jinyoung Han. 2024. HiQuE: Hierarchical Question Embedding Network for Multimodal Depression Detection. In *Proceedings of the 33rd ACM International*

*Corresponding Author.



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike International 4.0 License.

CIKM '24, October 21–25, 2024, Boise, ID, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0436-9/24/10.
<https://doi.org/10.1145/3627673.3679797>

Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA. ACM, New York, NY, USA, 11 pages.
<https://doi.org/10.1145/3627673.3679797>

1 Introduction

The diagnosis of depression in clinical settings often involves the use of interview-based instruments [66], in which mental health experts conduct clinical interviews with patients, assessing their symptoms [53, 66, 68, 89]. Due to the gradual and varied manifestation of depressive symptoms among individuals [13, 17, 18, 48], clinicians employ a structured interview process, which involves specific questionnaires and criteria, to detect a wide range of verbal and non-verbal symptoms of depression in patients' speech, behavior, facial expressions, and immediate responses during conversations [46, 71, 79].

Interview-based diagnostic methods have proven highly effective in interpreting patient responses [29, 53]. Mental health experts strategically incorporate follow-up questions in conjunction with primary questions to gather additional information from patients, thereby enhancing their understanding of the exhibited depressive symptoms [68]. In a case where a response from an initial follow-up question is insufficient for diagnosis, further follow-up questions can be employed to synthesize the patient's responses. By employing hierarchical questions during clinical interviews, clinicians can obtain a comprehensive understanding of depressive signals and the patient's overall condition, leading to improved diagnostic accuracy [17, 68].

Using clinical interview data, many scholars have proposed methods that can detect depression by analyzing revealed verbal (e.g., textual) or non-verbal (e.g., visual or acoustic) signals. Some studies have delved into visual cues, encompassing facial expressions and head poses [4, 5, 31, 51]. Additionally, a series of investigations has focused on acoustic and textual cues, with the goal of diagnosing depression based on linguistic patterns, vocal qualities, pitch, and loudness [2, 43, 63, 82, 91]. Nevertheless, these approaches have treated the entire input sequence as a singular entity, disregarding the structured nature of clinical interviews. Given the use of structured questionnaires and conversational exchanges during these interviews, considering the interview data as a single input sequence can be less effective for depression detection in

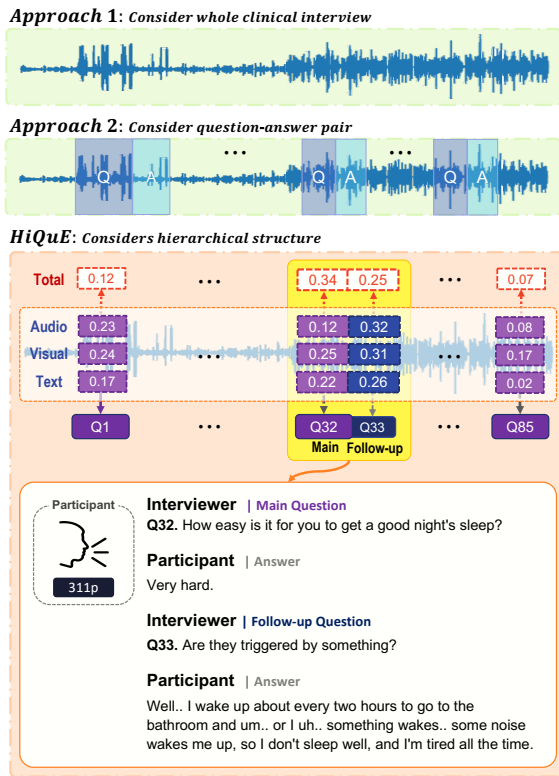


Figure 1: Previous research focused on learning the clinical interview sequences or their question-and-answer segments using a single modality. Our novel model, HiQuE, considers the hierarchy of questions incorporating multiple modalities to improve its effectiveness in depression detection. Numerical values in the boxes represent attention scores.

clinical settings [32, 87]. A few studies have sought to consider the conversations, including questions and answers during the clinical interviews [11, 45, 52, 80, 81, 87]. However, their focus has primarily been on analyzing questions and answers in an interview without specifically modeling the relationship between primary and follow-up questions based on question types, which can be crucial in modeling and analyzing the structure of a clinical interview. Besides, there has been a lack of analysis on the interaction among multi-modalities in analyzing structured clinical interviews [2, 19, 25, 37, 44, 59, 65, 87]; different modalities can be different cues depending on questions and answer types.

To address these challenges, we propose HiQuE (Hierarchical Question Embedding network), a novel depression detection framework that leverages the hierarchical relationship between primary and follow-up questions in a clinical interview. Inspired by the clinical interview strategy employed by medical professionals for diagnosing depression, HiQuE incorporates a hierarchical embedding structure and interview-specific attention modules. These modules enable HiQuE to comprehensively assess the mutual information between multiple modalities within interviews, replicating the diagnostic approach used by clinicians. As illustrated in Figure 1, the interview sequence is divided into primary questions and

their corresponding follow-up questions. Using the question-aware module, HiQuE calculates the significance of each question and effectively enhances the mutual information across modalities using cross-modal attention, resulting in accurate depression diagnosis. The contributions of this study can be summarized as follows:

- To the best of our knowledge, HiQuE is the first attempt that analyzes the significance of all questions posed by the interviewer by explicitly categorizing them as primary and follow-up questions, considering their order and relationship. Our publicly available code¹ encompasses both the hierarchical question embedding process and the HiQuE.
- This is the first interpretable multi-modal analysis conducted in a clinical interview context by analyzing both intra-modality and inter-modality attention scores. The quantitative evaluation of the interaction and importance of different modalities in depression detection provides deeper insights into the complex dynamics of clinical interviews.
- HiQuE achieves the state-of-the-art performance on the DAIC-WOZ dataset, among other multimodal emotion recognition models as well as prior depression detection models that utilized the DAIC-WOZ dataset. HiQuE also demonstrates superior adaptation to the E-DAIC-WOZ dataset, highlighting its generalizability to unseen question scenarios.

2 Related Work

2.1 Multimodal Expressions of Major Depressive Disorder in Clinical Interviews

Researchers have identified distinctive features across various modalities, including acoustic patterns, visual characteristics, and language usage in clinical interviews. For instance, individuals with depression often exhibit specific acoustic features, such as slower speaking rates, lower pitch ranges, and reduced loudness [6, 9, 20, 36, 75, 76, 79] as well as visual features, including discernible facial expressions characterized by sadness, minimal head movement [5], unstable facial expressions [14, 78], and irregular eye-gazing patterns [33, 40]. Moreover, they often reveal negative emotions in language, utilize a higher frequency of first-person pronouns, and exhibit intense focus on specific words [3, 61, 85]. These findings highlight the crucial role of incorporating multiple modalities for effective depression detection to develop a comprehensive understanding of an individual's symptoms [17, 18, 48].

2.2 Automatic Depression Detection

There have been a considerable number of proposals to detect depression by developing machine learning algorithms or natural language processing techniques [41, 47]. Initially, a substantial efforts were dedicated to extracting representative features [69, 70] and creating single-modality models for depression detection [30, 51, 77, 82, 83]. Furthermore, as Multimodal Sentiment Analysis (MSA) [49, 67] gained momentum with the recognition of various verbal and non-verbal symptoms of depression in psychological research, researchers made significant attempts to incorporate context-aware attention [10] and multimodal attention [26] to capture diverse

¹<https://github.com/JuHo-Jung/HiQuE>

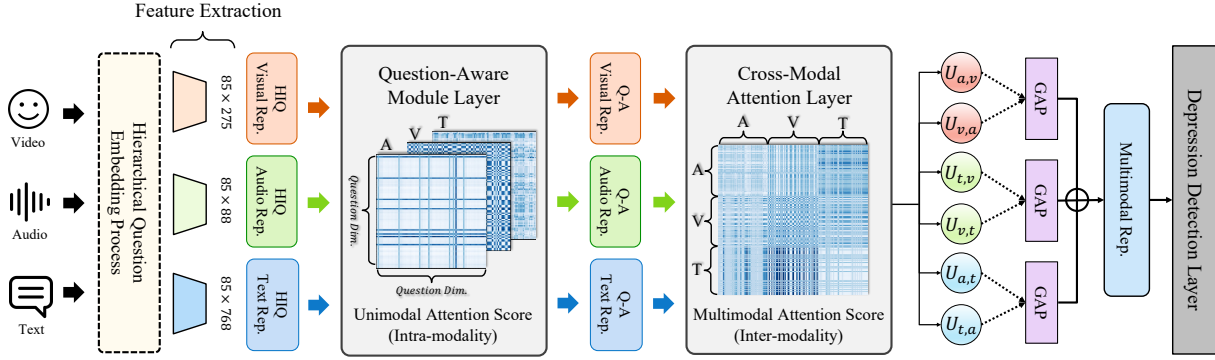


Figure 2: An overall architecture of the HiQuE’s multimodal depression detection process, where HIQ and Q-A indicate Hierarchical Question Embedded and Question-Aware, respectively.

information across multiple modalities [18, 23, 37, 62, 86, 88]. Recently, there have been attempts that analyze the word-sentence relations on interviewee’s answers [45, 59, 81, 87, 92] as well as the correlation between question-answer pairs [24, 52, 80], which can be cued in identifying depression. Unfortunately, no research exists yet that explains how an attention score of modality manifests in a specific question or how the degree of modality reflection changes with the sequence of primary and follow-up questions and answers, which can be crucial in modeling and analyzing the structure of a clinical interview. To bridge this gap, we introduce the first interpretable multimodal depression detection framework that leverages the hierarchical relationship between primary and follow-up questions in a clinical interview.

3 Clinical Interview Dataset

To train our proposed method for the depression detection task, we use the DAIC-WOZ dataset [28], which is a subset of the widely used dataset called Distress Analysis Interview Corpus (DAIC) [74]. The DAIC-WOZ dataset comprises clinical interviews conducted to diagnose psychological distress disorders. These interviews involve Wizard-of-Oz interactions, where an AI virtual interviewer named Ellie is controlled by a human interviewer located remotely. The dataset consists of speech samples from 189 participants, including audio/visual features, raw audio files, and interview transcripts. Following the prescribed guidelines, we split the dataset into 107 training samples, 35 validation samples, and 47 test samples.

3.1 Data Augmentation with Random Sampling

The DAIC-WOZ dataset suffers from a significant class imbalance, with a higher proportion of non-depression samples. Some prior studies addressed this issue by employing data augmentation techniques like random masking [7, 43, 65]. Inspired by these, we tripled the size of the depression dataset by randomly masking 10 out of 85 questions in each $85 \times N$ question-embedded interview sequence, aligning it with the size of the non-depression dataset during training. Specifically, we first segmented the interview sequences into question-answer (Q-A) pairs based on timestamps, starting from the interviewer’s question to the participant’s response. Then, we randomly masked ten Q-A pairs per interview, corresponding to the interviewer’s questions. Unused questions

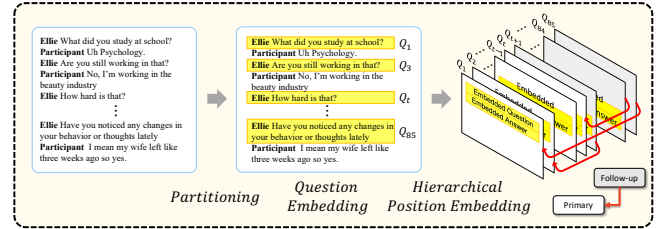


Figure 3: Hierarchical Question Embedding Process.

were replaced with zero vectors. This augmentation process was applied during only training.

4 Hierarchical Question Embedding Network

4.1 Problem Statement

Suppose we have a set of depression dataset $C = \{c_i\}_{i=1}^{|C|}$, where c_i contains the multi-modal inputs including audio, video, and text sequences; $X_a \in \mathbb{R}^{L_a \times d_a}$, $X_v \in \mathbb{R}^{L_v \times d_v}$, and $X_t \in \mathbb{R}^{L_t \times d_t}$, where L_m represents the sequence length and d indicates the feature dimension. Given the hierarchical structure of interview questions, we segment the interview sequence into question-answer pairs. Specifically, each input sequence is defined as $S = \{s_i\}_{i=1}^n$, where $s_i = (\text{question}_i, \text{answer}_i)$. We then annotate the input sequence S with corresponding hierarchical positions in the hierarchical question embedding process, denoted as $\hat{S} = \{(s_i, \text{pos}_i)\}_{i=1}^n$, where pos_i indicates the hierarchical position of the question_i . Finally, the proposed model predicts an individual c_i depression symptom $\hat{y} \in \{\text{normal}, \text{depression}\}$.

4.2 Overall Architecture

The proposed method, HiQuE, as shown in Figure 2, consists of three layers: (i) *Question-Aware Module*, (ii) *Cross-Modal Attention Layer*, and (iii) *Depression Detection Layer*. HiQuE categorizes interview sequences into main and follow-up questions using a hierarchical question embedding process. Audio, visual, and text features are extracted separately, and the *Question-Aware Module* generates Question-Aware representations for each feature. These are combined in the *Cross-Modal Attention Layer* to create a final multimodal representation, which the *Depression Detection Layer* uses to predict the presence of depression.

4.3 Hierarchical Question Embedding Process

As depicted in Figure 3, interviewer Ellie’s questions are categorized into 85 topics based on content, following Gong et al.’s approach [27]. Each question is associated with a specific topic code, such as labeling “How has seeing a therapist affected you?” as *therapist_affect* and “Where are you from originally?” as *origin*. These questions are further categorized into 66 primary and 19 follow-up questions based on content and order. For a complete list of the questions, please refer to Table 5. Finally, we systematically tag each question based on its hierarchical order, specifically when a follow-up question follows a primary question or when a follow-up question follows a previous follow-up question. For instance, where the question sequence is “What did you study at school?”, “Are you still working in that?”, and “How hard is that?”, the hierarchical order would be primary – follow-up – follow-up.

An overall process of hierarchical question embedding is depicted in Figure 3. Interview sequences are represented as unimodal raw sequences X_m , where m denotes *modality* $\in \{a, v, t\}$, respectively. Sequences X_m are partitioned into segments $S = \{s_i\}_{i=1}^n$ based on question and answer boundaries. Note that the number of segments, n , may vary for each sample due to differences in the type and number of questions employed during each interview. Then, each segment is split into a question segment and an answer segment; $S = \{(q_1, a_1) \cdots (q_i, a_i) \cdots (q_{85}, a_{85})\}$. After partitioning, segments are labeled with topic codes corresponding to each question and given hierarchical position embeddings based on their relationships. Specifically, as shown in Figure 3, we assign the previous question’s *Topic id* to the follow-up question. These hierarchical positions are incorporated into the representation before feeding them into the model. This embedding process ensures uniform vector shapes by replacing unused questions with zero vectors, resulting in 85-dimensional representations for all samples.

4.4 Feature Extraction

4.4.1 Audio Feature: For audio feature extraction, we utilize the open-Source Media Interpretation by Large feature-space Extraction (openSMILE) [22], along with the extended Geneva Minimalistic Acoustic Parameter Set [21]. These features encompass 88 functionals, including loudness, MFCCs, and other characteristics that aid in discerning emotions in speech. Consequently, each interviewee’s audio features are represented as 85×88 -dimensional vectors, where 85 denotes the question embedding dimension. These audio features are then processed using a transformer encoder.

4.4.2 Visual Feature: Due to privacy concerns, the dataset only offers visual features extracted via the Constrained Local Neural Fields (CLNF) algorithm [8], a widely-used approach for facial landmark localization and face recognition. To address the variation in interview duration for each answer, we first extract 68 facial landmarks from each frame (at a rate of 1 frame per second) within each segment, considering their respective x and y coordinates. We then compute mean and variance vectors within each segment and concatenate the x and y coordinates. This results in 85×272 -dimensional vectors per participant, with zero vectors used for segments where a face is not detected.

4.4.3 Text Feature: For text feature extraction, we segment the interview transcripts into sections corresponding to individual answers for each question. We next leverage the pre-trained RoBERTa [42] to generate text features from each answer segment. Given the RoBERTa’s strength in robustly capturing contextual information and semantic nuances in various NLP task [38], it demonstrated superior performance compared to other embedding methods and large language models (LLMs) as shown in Section 5.3. We extract features from the final layer, focusing on the [CLS] token, resulting in an 85×768 -dimensional vector for each answer, where 85 represents the dimensionality of the question embedding.

4.5 Question-Aware Module Layer

In Figure 2, a transformer encoder with h multi-heads is utilized to capture attention between questionnaire responses. Initially, a stack of 1-dimensional convolutional layers is applied to process local information, converting varying shapes of HIQ Visual Rep. (85×275), HIQ Visual Rep. (85×88), and HIQ Visual Rep. (85×768) into uniform shapes of 85×4 denoted as U_m , $m \in t, a, v$.

Subsequently, the question-aware self-attention mechanism guides the transformer encoder to focus on important segments and relationships among the question-embedded sequences. Given that each representation is embedded based on 85 questions, self-attention allows the model to focus on important questions within the question-embedded representation. As a result, this particular attention mechanism enables HiQuE to extract meaningful information, represented as Q-A M Rep; $M \in \{Audio, Visual, Text\}$, in the form of 85×85 matrices, for depression detection from each question. We analyze this unimodal attention score to identify the significant components of intra-modality. The same input is employed for self-attention as query (Q), key (K), and value (V) in the following equations:

$$MultiHead(Q, K, V) = Concatenate(head_1, \dots, head_h) \quad (1)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

$$Q-A \ m \ Rep. = MultiHead(U_m, U_m, U_m) + U_m \quad (3)$$

4.6 Cross-Modal Attention Layer

In Figure 2, the multimodal transformer encoder with h multi-heads integrates information from two modalities using a cross-attention mechanism [34]. This mechanism allows the model to discern crucial relationships between m_1 and m_2 modalities, with m_1 serving as the source (query) and m_2 as the target (key and value). Furthermore, since the information in the two modalities differs, we conduct bidirectional cross-attention between m_1 and m_2 (i.e., audio-visual, visual-text, and text-audio) to allow the model to learn relevant information across modalities as follows:

$$U_{m_1, m_2} = MultiHead(U_{m_1}, U_{m_2}, U_{m_2}) + U_{m_1} \quad (4)$$

$$U_{m_2, m_1} = MultiHead(U_{m_2}, U_{m_1}, U_{m_1}) + U_{m_2} \quad (5)$$

Given that the input to the cross-modal attention layer is U_m , $m \in t, a, v$ from the question-aware module layer, each input representation has a shape of 85×85 . This allows us to analyze the multimodal attention score to identify significant components between different modalities.

Table 1: Performance comparisons on the DAIC-WOZ dataset among twelve state-of-the-art baseline models and the proposed model. (WA*: Weighted Average)

Method	Approach	Precision	Recall	F1-Score	WA* Prec. (↑)	WA* Rec. (↑)	WA* F1 (↑)	G-Mean (↑)
<i>TFN</i> [90]	Modality-Aware	0.67	0.73	0.68	0.84	0.78	0.81	0.699
<i>BiLSTM-1DCNN</i> [41]	Modality-Aware	0.65	0.61	0.62	0.77	0.71	0.73	0.630
<i>Mult</i> [73]	Modality-Aware	0.73	0.74	0.74	0.81	0.77	0.77	0.735
<i>MISA</i> [35]	Modality-Aware	0.74	0.77	0.74	0.86	0.77	0.79	0.755
<i>D-vlog</i> [88]	Modality-Aware	0.73	0.72	0.73	0.82	0.76	0.77	0.725
<i>bc-LSTM</i> [55]	Context-Aware	0.59	0.60	0.59	0.77	0.69	0.72	0.595
<i>Emotion Recognition</i> [64]	Context-Aware	0.65	0.69	0.66	0.69	0.70	0.71	0.670
<i>Sequence Modeling</i> [2]	Context-Aware	0.67	0.71	0.70	0.85	0.73	0.77	0.690
<i>Topic Modeling</i> [27]	Context-Aware	0.63	0.60	0.62	0.81	0.71	0.74	0.615
<i>Context Aware</i> [37]	Context-Aware	0.71	0.71	0.71	0.85	0.73	0.77	0.710
<i>Speechformer</i> [11]	Context-Aware	0.70	0.72	0.70	0.78	0.76	0.76	0.710
<i>GRU/BiLSTM-based</i> [65]	Context-Aware	0.75	0.78	0.75	0.86	0.77	0.80	0.765
<i>HiQuE</i>	Modality + Context	0.78	0.80	0.79	0.85	0.80	0.82	0.790

4.7 Depression Detection Layer

In the last stage, the *audio-visual*, *visual-text*, and *text-audio* cross-modal representations are transformed into a final multimodal representation after layer normalization, concatenation, and GAP (global average pooling), as follows:

$$\tilde{U} = \sum \text{GAP}(U_{m_1, m_2} \oplus U_{m_2, m_1}) \quad (6)$$

Finally, multimodal representation is fed into HiQuE’s depression detection layer to detect depression as follows:

$$\hat{Y} = \text{softmax}(\text{HiQuE}(\tilde{U})) \quad (7)$$

where the HiQuE prediction layer comprises a fully connected layer and a dropout layer. Since the depression detection task can be defined as a binary classification problem, we employed the cross entropy as the loss function as follows:

$$\text{Loss} = -\frac{1}{b} \sum_{i=1}^b [y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)] \quad (8)$$

where b represents the batch size, i is an index representing each example within the batch, y_i is the actual label where 0 represents normal and 1 represents depression, and \hat{y}_i is the softmax function that represents the model’s prediction or probability.

5 Experiments

We use Tensorflow [1] to implement the proposed model. The dropout rate, batch size, epochs, and learning rate were set to 0.5, 8, 100, and 0.0002, respectively. The maximum sequence length was set to 85 since all sequences are embedded into 85 questions. All weights are randomly initialized in both *HiQuE* and baselines.

5.1 Baseline Methods

To evaluate the overall performance of the proposed model, we compare its performance against five state-of-the-art multimodal models for depression detection and emotion recognition as follows: (i) Tensor Fusion Network (*TFN*) [90], (ii) bidirectional LSTM / 1D CNN-based model (*BiLSTM-1DCNN*) [41], (iii) Multimodal Transformer (*Mult*) [73], (iv) *MISA* [35], and (v) *D-vlog* [88]. Since these models were specifically designed to analyze multimodal fusion methods, we have categorized them as “*Modality-Aware*”.

We further utilize seven context-aware multimodal models for depression detection and emotion recognition to compare the analysis of the hierarchical structure of clinical interviews: (i) bidirectional contextual LSTM (*bc-LSTM*) [55], (ii) *Emotion Recognition* [64],

Table 2: Performance on different text embeddings.

Text Embedding	Precision	Recall	F1-Score
BART [39]	0.75	0.76	0.74
GloVe [54]	0.72	0.73	0.72
BERT [15]	0.76	0.77	0.76
CodeLlama [60]	0.78	0.75	0.76
Llama2 [72]	0.77	0.77	0.77
GPT-2 [57]	0.76	0.78	0.77
RoBERTa [42] (Ours)	0.78	0.80	0.79

(iii) *Sequence Modeling* [2], (iv) *Topic Modeling* [27], (v) Context-aware deep learning (*Context-Aware*) [37], (vi) *Speechformer* [11] and (vii) *GRU/BiLSTM-based* [65]. As these methods consider the context of interview questions and answers or focus on the topics of questions and the timing of their appearance during the interview, we have categorized them as “*Context-Aware*”.

Note that we extract multimodal features from the entire interview sequence to train the five modality-aware methods and seven context-aware methods. All models were trained on the same data partition to ensure fairness and evaluated using the hyperparameters that showed the best performance.

5.2 Experimental Results

To provide a comprehensive assessment of the models’ performance, particularly in the context of an imbalanced dataset (i.e., DAIC-WOZ), we report experimental results with various metrics including the weighted average and geometric mean scores (G-mean score). Table 1 shows the Macro Average precision / recall / F1-score, Weighted Average precision / recall / F1-score, and G-mean score of the baseline models and the proposed model, respectively.

As shown in Table 1, *HiQuE* achieves the best depression detection with a macro average F1-score of 0.79, a weighted average F1-score of 0.82, and a G-mean score of 0.790. As macro-average treats each class equally, while weighted-average gives weight based on class size, the result that *HiQuE* excels in both metrics highlights *HiQuE*’s robustness and effectiveness against an imbalanced dataset, showcasing its ability to capture distinct depression indicators.

Among the baseline models, *GRU/BiLSTM-based* [65] achieves the highest performance with a macro average F1-score of 0.75, weighted average precision of 0.86, and G-mean score of 0.765. This underscores the effectiveness of analyzing speech characteristics

Table 3: Validating the generalizability of *HiQuE* on the two datasets (E-DAIC-WOZ and MIT Interview) using multimodal inputs (audio, video, and text). Tasks include binary classification for Depression Detection, Stress Level Prediction, and Job Interview Performance, respectively.

Methods	E-DAIC-WOZ [58]			MIT Interview dataset [50]					
	Depression Detection			Stress Level Prediction			Job Interview Performance Prediction		
	Overall			Overall			Overall		
	Pre.	Rec.	F1.	Pre.	Rec.	F1.	Pre.	Rec.	F1.
<i>GRU/BiLSTM-based</i> [65]	0.67	0.63	0.65	0.70	0.76	0.73	0.73	0.75	0.74
<i>D-vlog</i> [88]	0.65	0.69	0.67	0.71	0.73	0.72	0.71	0.72	0.72
<i>MISA</i> [35]	0.62	0.64	0.63	0.69	0.68	0.68	0.69	0.69	0.69
<i>MuT</i> [35]	0.64	0.65	0.64	0.70	0.73	0.72	0.70	0.72	0.71
<i>HiQuE</i>	0.71	0.70	0.70	0.75	0.81	0.78	0.76	0.79	0.77

and linguistic content within individual utterances for depression diagnosis within interview sequences. Furthermore, *MISA* demonstrates the second-highest performance among the baselines with a macro average F1-score of 0.74, weighted average precision of 0.86, and G-mean score of 0.755. This suggests that incorporating two subspaces (modality-invariant and modality-specific) allows the model for a comprehensive understanding of multimodal data, which suggests that for accurate multimodal depression detection, it is essential to capture both inter- and intra-representations between modalities. We also find that *MuT* and *D-vlog* exhibit promising performance at 0.74 and 0.73 of the macro average F1-score, respectively. This suggests that employing a cross-attention mechanism to learn the relationship between multiple modalities helps the model learn important signals for depression detection.

5.3 Text Embedding Performance Comparison

We chose to use the pre-trained RoBERTa [42] as an encoder due to its higher performance as shown in in Table 2, in comparison with other popular embedding techniques and large language models (LLMs). The high performance of RoBERTa is due to its robust representations and comprehensive contextual understanding. LLMs also showed a comparable performance as shown in Table 2, but we decided not to use them due to practical challenges related to privacy and stability, particularly in mental health applications.

5.4 Generalization to Unseen Questions

To assess *HiQuE*'s generalizability, we further utilized the E-DAIC-WOZ [58] and MIT Interview dataset [50]. The E-DAIC-WOZ [58] comprises audio-visual recordings of semi-clinical interviews conducted in English, featuring numerous questions absent in the DAIC-WOZ dataset. However, it does not provide the transcript of interviewer's questions, making it difficult to determine the specific questions asked. The MIT Interview dataset [50] includes 138 interview videos of internship-seeking students from MIT, featuring facial expressions, language use, and prosodic cues. Moreover, it provides ground truth labels for stress level and job interview performance, rated by nine independent judges. This dataset encompasses multimodal features influencing mental states during job interviews [50].

We adapt our model to these datasets by extracting text from the audio using the whisper [56] and mapping unseen questions to the predefined list (Table 5) based on the BERT-score² similarity.

²<https://huggingface.co/spaces/evaluate-metric/bertscore>

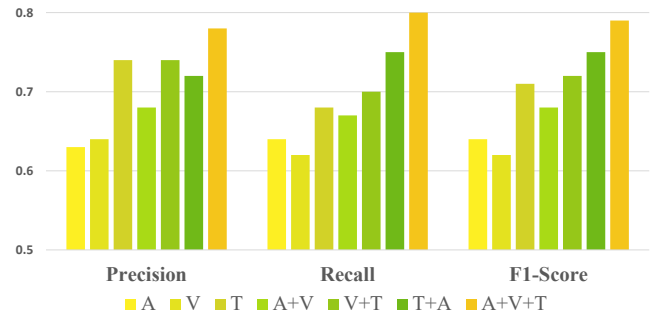


Figure 4: Performance comparisons between unimodal and multimodal depression detection models. A, V, and T denote audio, visual, and text modality, respectively. The X-axis indicates the macro average precision, recall, and F1 score, while the Y-axis represents the corresponding scores.

If a question's similarity falls below the average, we consider it as a new question and add it to the list. Baselines adopted the same encoder as *HiQuE* but without hierarchical question embedding. Table 3 shows that *HiQuE* outperforms the baselines across both datasets for the three different tasks: Depression Detection, Stress Level Prediction, and Job Interview Performance Prediction. This underscores *HiQuE*'s effectiveness in detecting depression cues from clinical interviews, even with non-predefined questions. More importantly, the experimental result that *HiQuE* has shown promising performance not only in clinical interviews but also in job interviews highlights its usability in various real-world interview scenarios.

6 Analysis

6.1 Analysis on Different Modalities

To analyze the importance of each modality (i.e., audio, visual, and text) for detecting depression, we compare the performance of models that are trained with different sets of modalities. For the **unimodal models** (i.e., A, V, T), we first simply utilize a hierarchical question embedding process followed by a question-aware module layer for each input modality. We then add global average pooling and fully connected layers with softmax activation function to generate predicted labels (i.e., depressed or not). As shown in Figure 4, the model trained with text achieves the highest performance (0.71 of macro average F1-score) among the unimodal models. This implies that the text modality contains the most useful information in depression detection, which can be linked to the results of the prior

Table 4: The results of the ablation study on hierarchical question embedding process and model layers as illustrated in Figure 2, along with augmentation methods with random sampling as described in Section 3.1. The term “Q-A Module” denotes the Question-Aware Module Layer, while “C-M Attention” represents the Cross-Modal Attention Layer, which is examined as part of the model components. “Q.E.” and “H.Q.E.” stand for Question Embedding and Hierarchical Question Embedding, respectively, and “Aug.” refers to the Augmentation.

Ablation Settings					Precision	Recall	F1-Score	WA* F1 (↑)
Question Embedding		Model Layer		Aug.				
Q.E.	H.Q.E.	Q-A Module	C-M Attention					
✗	✗	✓	✓	✓	0.74	0.73	0.73	0.75
✓	✗	✓	✓	✓	0.75	0.76	0.75	0.77
✓	✓	✗	✗	✓	0.73	0.72	0.72	0.74
✓	✓	✓	✗	✓	0.73	0.75	0.74	0.77
✓	✓	✗	✓	✓	0.75	0.76	0.76	0.79
✓	✓	✓	✓	✗	0.76	0.77	0.76	0.80
✓	✓	✓	✓	✓	0.78	0.80	0.79	0.82

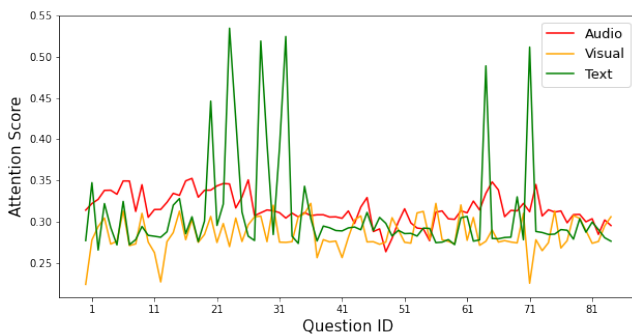


Figure 5: Distributions of attention scores across different modalities in each question, with the X-axis representing the Question ID from the first question (Q_1) to the last question (Q_{85}), and the Y-axis indicating the attention score.

studies [12, 16, 59, 84]. For **bimodal models** (i.e., A+V, V+T, A+T), we first fuse two unimodal encoders via a cross-modal attention layer. We then add the same depression detection layer as unimodal models. Since the text feature contains the most useful information (i.e., A+T, V+T) show higher performance than the model trained without text modality (i.e., A+V). Also, we find that considering **all modalities** (i.e., A+V+T) significantly improves performance. This reveals that learning both verbal and non-verbal signals, as well as their relationships, is an effective way for depression detection.

6.2 Intermodal Interaction Analysis

By examining the attention score distributions across different modalities, as depicted in Figure 5, we highlight the significance of each modality in depression detection. Notably, questions directly related to emotions or past experiences, such as “Tell me about an event or something that you wish you could erase from your memory?” or “Tell me about the last time you felt really happy?”, had a significant impact on both audio and visual modalities, while those related to a current emotional state or past depression diagnosis, such as “Have you been diagnosed with depression?” or “How have you been feeling lately?”, had the highest impact on text.

We also explore the impact of individual modalities (i.e., audio, visual, and text) when the model fails to make accurate predictions. Depressed patients are often misclassified as normal when interviewees exhibit cheerful tones or frequent laughter (resulting in

high audio and visual attention scores, Figure 5), or when there are no clear indicators of depression during the interview. Conversely, our model tends to misclassify normal as depression when negative words are frequently used, particularly when participants express recent feelings of anxiety and depression. In these cases, as highlighted in Figure 5, the text attention score predominantly influences the incorrect predictions.

6.3 Ablation Study

6.3.1 Hierarchical Question Embedding To highlight the benefits of our proposed hierarchical question embedding process, we conducted an ablation study with three distinct cases: *Non-Question Embedding* (N.Q.E), only *Question Embedding* (Q.E.), and *Hierarchical Question Embedding* (H.Q.E.), as shown in Table 4. In the case of *N.Q.E*, the entire interview sequence is treated as a single sequence for the depression detection model. Specifically, in this case, the input sequences are cropped from the beginning of the utterance to the end of the interview conversation. When only *Q.E* is applied, the interview is segmented into question-answer pairs. Notably, this procedure only divides the sequence into questions and aligns them with the respective question topics without incorporating hierarchical positional embedding. The improvement presented in Table 4 highlights that, for depression detection with clinical interviews, extracting useful information based on a question-driven approach is more effective than considering the entire interview sequence as a single sequence. In the last scenario *H.Q.E*, hierarchical position embedding is introduced following the question embedding procedure. To elaborate, after dividing the interview into question-answer pairs using question embedding, a hierarchical relationship (primary or follow-up) among the questions is tagged through hierarchical position embedding. As shown in Table 4, our proposed hierarchical question embedding process effectively forces the model to capture hierarchical relationships and the importance of the questions.

6.3.2 Model Components To assess the effectiveness of each layer in the HiQuE, we conducted an ablation study on the *Question-Aware Module Layer* and the *Cross-Modal Attention Layer*. As shown in Table 4, without “*Q-A Module Layer*”, each audio, visual, and text representation undergoes hierarchical question embedding and feature extraction before entering the *Cross-Modal Attention Layer*, which incorporates bidirectional cross-attention (e.g.,

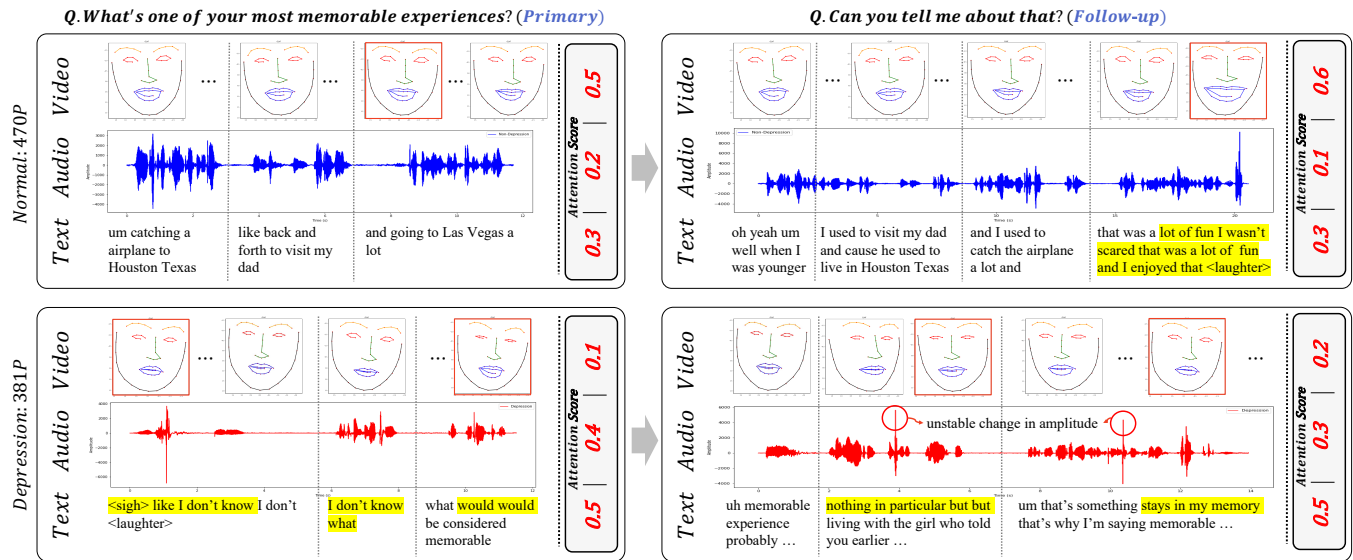


Figure 6: A case analysis of a depression case (381P) and a non-depression case (470P). Specific audio, video, and text responses are provided in response to the primary question, “What’s one of your most memorable experiences?” and the follow-up question, “Can you tell me about that?”. The highlighted sections in the figures (i.e., red rectangles in facial features or yellow highlights in the textual features) indicate distinct characteristics between a depression case and a non-depression case.

$U_{a,v} - U_{v,a}$) as illustrated in Figure 2. Given that the *Question-Aware Module Layer* assesses the relevance, importance, and mutual influence of the 85 embedded questions for each modality, its absence results in a performance degradation of the model. Without “*C-M Attention Layer*”, the HIQ Audio Rep., HIQ Visual Rep., and HIQ Text Rep. go through the *Question-Aware Module Layer*, and are then concatenated before entering the *Depression Detection Layer*. Since the *Cross-Modal Attention Layer* computes relevant information from different modalities, the result highlights the effectiveness of considering information from both different modalities for accurate depression detection. Interestingly, as depicted in Table 4, the setting without “*Q-A Module Layer*” achieves a higher macro average F1-score and weighted average F1-score compared to the setting without “*C-M Attention Layer*”. This reveals that in detecting depression, it is more important to learn relevant information and interactions between modalities than to analyze the relationships and importance of each question. Furthermore, Table 4 also presents the performances of our proposed model with and without data augmentation. The results confirm that data augmentation enhances performance by balancing the sizes of depression and non-depression cases in the training set.

7 Case Study

In this section, we present a case study on samples from our test set to assess the effectiveness of the decision-making process of *HiQuE*. Specifically, we examine the verbal (i.e., text) and non-verbal (i.e., audio and video) signals for the two cases: a depressed individual (381P) and a non-depressed individual (470P). Our analysis focuses on the distinct attributes of audio, text, and visual attention scores for each individual. For a fair comparison, we apply a normalization technique to the amplitude and time of the audio waves, allowing for unbiased and consistent analysis and comparison.

Figure 6 showcases how the model integrates text, audio, and visual features during the decision-making process for each questionnaire response. In the case of the primary question “What’s one of your most memorable experiences?”, we observe that the depressed individual faces difficulties in providing a prompt response. He/she exhibits hesitation while reflecting on memorable experiences and ultimately struggles to provide a specific answer. In contrast, the non-depressed individual is more likely to respond immediately and accurately.

By analyzing audio, visual, and text attention scores to the follow-up question “Can you tell me about that?”, we observe the comprehensive exploration and understanding of various responses exhibited by *HiQuE* in detecting depression. In the case of a non-depressed individual (470P), detailed explanations, expressions of excitement, and smiling faces are evident in the answer to the follow-up question. Note that *HiQuE* also gives the highest attention score 0.6 to visual features. Furthermore, apart from <laughter>, the audio waves display symmetrical patterns without irregular fluctuations, indicating a more wide range of tones and amplitudes. On the other hand, the depressed individual (381P) encounters difficulties recalling memorable experiences when responding to the follow-up question. Instead of positive recollections, this individual shares memories of regrettable past incidents. By examining the audio wave of the depressed individual, we observe unstable fluctuations in amplitude while his/her facial expressions remain neutral. For this reason, *HiQuE* assigns the highest attention score of 0.5 to text features, followed by attention scores of 0.3–0.4 for audio features.

Our analysis of the attention scores for each modality during the model’s diagnostic process demonstrates that *HiQuE* effectively incorporates the interview structure through its hierarchical question embedding layer. The case study provides further evidence

Table 5: Question list in the DAIC-Woz dataset. All questions were categorized into 66 main questions and 19 follow-up questions using hierarchical question mapping.

Ind.	Question	Type
(1)	how has seeing a therapist affected you	Primary
(2)	tell me about the last time you felt really happy	Primary
(3)	where are you from originally	Primary
(4)	when was the last time you argued with someone and what was it about	Primary
(5)	what advice would you give to yourself ten or twenty years ago	Primary
(6)	how are you at controlling your temper	Primary
(7)	what are some things you really like about l_a	Primary
(8)	what are you most proud of in your life	Primary
(9)	who's someone that's been a positive influence in your life	Primary
(10)	how would your best friend describe you	Primary
(11)	what are some things you don't really like about l_a	Primary
(12)	what did you study at school	Primary
(13)	is there anything you regret	Primary
(14)	what's your dream job	Primary
(15)	what do you enjoy about traveling	Primary
(16)	what are you like when you don't sleep well	Primary
(17)	what's one of your most memorable experiences	Primary
(18)	tell me about the hardest decision you've ever had to make	Primary
(19)	what are some things you like to do for fun	Primary
(20)	tell me about a situation that you wish you had handled differently	Primary
(21)	tell me about an event or something that you wish you could erase from your memory	Primary
(22)	why did you move to l_a	Primary
(23)	what are some things you wish you could change about yourself	Primary
(24)	what would you say are some of your best qualities	Primary
(25)	how often do you go back to your home town	Primary
(26)	how long ago were you diagnosed	Primary
(27)	what's something you feel guilty about	Primary
(28)	when did you move to l_a	Primary
(29)	how easy was it for you to get used to living in l_a	Primary
(30)	when was the last time you felt really happy	Primary
(31)	what's the hardest thing about being a parent	Primary
(32)	do you still go to therapy now	Primary
(33)	do you travel a lot	Primary
(34)	have you ever served in the military	Primary
(35)	when was the last time that happened	Primary
(36)	what's the best thing about being a parent	Primary
(37)	what are some things that make you really mad	Primary
(38)	do you find it easy to be a parent	Primary
(39)	what do you do now	Primary
(40)	what were your symptoms	Primary
(41)	tell me how you spend your ideal weekend	Primary
(42)	what do you do when you are annoyed	Primary
(43)	tell me about your kids	Primary
(44)	tell me about a time when someone made you feel really badly about yourself	Primary
(45)	what are some ways that you're different as a parent than your parents	Primary
(46)	what do you think of today's kids	Primary
(47)	do you feel down	Primary
(48)	how do you like your living situation	Primary
(49)	how are you doing today	Primary
(50)	do you have roommates	Primary
(51)	do you think that maybe you're being a little hard on yourself	Primary
(52)	do you have disturbing thoughts	Primary
(53)	where do you live	Primary
(54)	what did you do after the military	Primary
(55)	did you ever see combat	Primary
(56)	why don't we talk about that later	Primary
(57)	how did serving in the military change you	Primary
(58)	have you noticed any changes in your behavior or thoughts lately	Primary
(59)	have you been diagnosed with depression	Primary
(60)	how easy is it for you to get a good night sleep	Primary
(61)	how close are you to your family	Primary
(62)	how have you been feeling lately	Primary
(63)	do you consider yourself an introvert	Primary
(64)	have you ever been diagnosed with p_t_s_d	Primary
(65)	do you feel like therapy is useful	Primary
(66)	what do you do to relax	Primary
(67)	can you tell me about that	Follow-up
(68)	why	Follow-up
(69)	how hard is that	Follow-up
(70)	what made you decide to do that	Follow-up
(71)	are you still doing that	Follow-up
(72)	what got you to seek help	Follow-up
(73)	how do you cope with them	Follow-up
(74)	how does it compare to l_a	Follow-up
(75)	are you okay with this	Follow-up
(76)	are they triggered by something	Follow-up
(77)	are you happy you did that	Follow-up
(78)	could you have done anything to avoid it	Follow-up
(79)	has that gotten you in trouble	Follow-up
(80)	how do you know them	Follow-up
(81)	do you feel that way often	Follow-up
(82)	did you think you had a problem before you found out	Follow-up
(83)	why did you stop	Follow-up
(84)	what's it like for you living with them	Follow-up
(85)	can you give me an example of that	Follow-up

that *HiQuE* successfully captures the sequential information of all questions and maximizes the mutual information between modalities by leveraging the question-aware module and the cross-modal attention layer.

8 Conclusion

In this paper, we presented HiQuE, a novel hierarchical question embedding model for multimodal depression detection. HiQuE efficiently captures the hierarchical structure of questions in clinical interviews and explores the correlations between different modalities to extract valuable information for depression detection. Through a comprehensive case study, we confirmed that the HiQuE focuses on questions specifically related to depression and makes its final decision by utilizing attention scores. This approach allows the model to mimic the expertise of clinical professionals during clinical interviews, where the interaction of questionnaire responses plays a crucial role. Given HiQuE's demonstrated generalizability to unseen questions, future plans involve extending its applicability to additional speech-related tasks and exploring the advantages of hierarchical question embedding further.

Acknowledgments

This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program (RS-2024-00425354) and the Graduate School of Metaverse Convergence support program (IITP-2024-RS-2023-00254129) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

A List of the 85 Questions in DAIC-Woz Dataset

We present the comprehensive list of 85 questions employed in the DAIC-WOZ dataset. Each question posed by the interviewer (Ellie) was mapped to the relevant topics using the question's topic codes [27]. We augmented the list with additional questions from the interviewer (Ellie) and rectified any inaccuracies in the existing questions. Furthermore, we organized all the questions into a hierarchical structure, comprising 66 primary questions and 19 follow-up questions, determined by their content and the order in which they were posed.

References

- [1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: a system for large-scale machine learning. In *Osdi*, Vol. 16. Savannah, GA, USA, 265–283.
- [2] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting Depression with Audio/Text Sequence Modeling of Interviews. In *Interspeech*. 1716–1720.
- [3] Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science* 6, 4 (2018), 529–542.
- [4] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Matthew Hyett, Gordon Parker, and Michael Breakspear. 2016. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing* 9, 4 (2016), 478–490.
- [5] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parker, and Michael Breakspear. 2013. Head pose and movement analysis as an indicator of depression. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 283–288.
- [6] Murray Alpert, Enrique R Pouget, and Raul R Silva. 2001. Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders* 66, 1 (2001), 59–69.
- [7] Andrew Bailey and Mark D Plumbley. 2021. Gender bias in depression detection using audio features. In *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 596–600.
- [8] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops*. 354–361.

- [9] Michael Cannizzaro, Brian Harel, Nicole Reilly, Phillip Chappell, and Peter J Snyder. 2004. Voice acoustical measurement of the severity of major depression. *Brain and cognition* 56, 1 (2004), 30–35.
- [10] Dushyant Singh Chauhan, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5647–5657.
- [11] Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2022. Speechformer: A hierarchical efficient framework incorporating the characteristics of speech. *arXiv preprint arXiv:2203.03812* (2022).
- [12] Raymond Chiong, Gregorius Satia Budhi, Sandeep Dhakal, and Fabian Chiong. 2021. A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine* 135 (2021), 104499.
- [13] Nicholas Cummins, Jyoti Joshi, Abhinav Dhall, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps. 2013. Diagnosis of depression by behavioural signals: a multimodal approach. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. 11–20.
- [14] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech communication* 71 (2015), 10–49.
- [15] 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* (2018).
- [16] Shubham Dham, Anirudh Sharma, and Abhinav Dhall. 2017. Depression scale recognition from audio, visual and text analysis. *arXiv preprint arXiv:1709.05865* (2017).
- [17] Hamdi Dibeklioglu, Zakia Hammal, and Jeffrey F Cohn. 2017. Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE journal of biomedical and health informatics* 22, 2 (2017), 525–536.
- [18] Hamdi Dibeklioglu, Zakia Hammal, Ying Yang, and Jeffrey F Cohn. 2015. Multimodal detection of depression in clinical interviews. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 307–310.
- [19] Yizhuo Dong and Xinyu Yang. 2021. A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing* 441 (2021), 279–290.
- [20] Heiner Ellgring and Klaus R Scherer. 1996. Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior* 20 (1996), 83–110.
- [21] 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 minimalist acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.
- [22] 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*. 1459–1462.
- [23] Ming Fang, Siyu Peng, Yujia Liang, Chih-Cheng Hung, and Shuhua Liu. 2023. A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control* 82 (2023), 104561.
- [24] Ricardo Flores, ML Tlachac, Erial Toto, and Elke Rundensteiner. 2022. Transfer learning for depression screening from follow-up clinical interview questions. In *Deep Learning Applications, Volume 4*. Springer, 53–78.
- [25] Ricardo Flores, ML Tlachac, Erial Toto, and Elke A Rundensteiner. 2021. Depression screening using deep learning on follow-up questions in clinical interviews. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 595–600.
- [26] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Contextual inter-modal attention for multi-modal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing*. 3454–3466.
- [27] 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*. 69–76.
- [28] Jonathan Gratch, Ron Artstein, Gale 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*. Technical Report. University of Southern California Los Angeles.
- [29] Jenny Guidi, Giovanni A Fava, Per Bech, and Eugene Paykel. 2010. The Clinical Interview for Depression: a comprehensive review of studies and clinimetric properties. *Psychotherapy and Psychosomatics* 80, 1 (2010), 10–27.
- [30] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100* (2020).
- [31] Yanrong Guo, Chenyang Zhu, Shijie Hao, and Richang Hong. 2022. Automatic depression detection via learning and fusing features from visual cues. *IEEE Transactions on Computational Social Systems* (2022).
- [32] Shan Guohou, Zhou Lina, and Zhang Dongsong. 2020. What reveals about depression level? The role of multimodal features at the level of interview questions. *Information & Management* 57, 7 (2020), 103349.
- [33] Rahul Gupta, Nikolaos Malandrakis, Bo Xiao, Tanaya Guha, Maarten Van Segbroeck, Matthew Black, Alexandros Potamianos, and Shrikanth Narayanan. 2014. Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*. 33–40.
- [34] Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 12972–12980.
- [35] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1122–1131.
- [36] Emil Kraepelin. 1921. *Manic-depressive insanity and paranoia*. E. & S. Livingstone.
- [37] Genevieve Lam, Huang Dongyan, and Weisi Lin. 2019. Context-aware deep learning for multi-modal depression detection. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 3946–3950.
- [38] Daeun Lee, Sejung Son, Hyolim Jeon, Seungbae Kim, and Jinyoung Han. 2023. Learning Co-Speech Gesture for Multimodal Aphasia Type Detection. *arXiv preprint arXiv:2310.11710* (2023).
- [39] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).
- [40] Yu Li, Yangyang Xu, Mengqing Xia, Tianhong Zhang, Junjie Wang, Xu Liu, Yongguang He, and Jijun Wang. 2016. Eye movement indices in the study of depressive disorder. *Shanghai Archives of Psychiatry* 28, 6 (2016), 326.
- [41] Lin Lin, Xuri Chen, Ying Shen, and Lin Zhang. 2020. Towards automatic depression detection: A BiLSTM/1D CNN-based model. *Applied Sciences* 10, 23 (2020), 8701.
- [42] 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* (2019).
- [43] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 35–42.
- [44] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 6818–6825.
- [45] Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn W. Schuller. 2019. A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews. In *Interspeech*. 221–225. <https://doi.org/10.21437/Interspeech.2019-2036>
- [46] Ramin Mojtabai. 2013. Clinician-identified depression in community settings: concordance with structured-interview diagnoses. *Psychotherapy and psychosomatics* 82, 3 (2013), 161–169.
- [47] Michelle Morales, Stefan Scherer, and Rivka Levitan. 2017. A cross-modal review of indicators for depression detection systems. In *Proceedings of the fourth workshop on computational linguistics and clinical psychology—From linguistic signal to clinical reality*. 1–12.
- [48] Michelle Renee Morales and Rivka Levitan. 2016. Speech vs. text: A comparative analysis of features for depression detection systems. In *2016 IEEE spoken language technology workshop (SLT)*. IEEE, 136–143.
- [49] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*. 169–176.
- [50] Iftekhar Naim, Md Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2016. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing* 9, 2 (2016), 191–204.
- [51] Sana A Nasser, Ivan A Hashim, and Wisam H Ali. 2020. A review on depression detection and diagnoses based on visual facial cues. In *2020 3rd International Conference on Engineering Technology and its Applications (IICETA)*. IEEE, 35–40.
- [52] 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)*. IEEE, 4235–4239.
- [53] ES Paykel. 1985. The Clinical Interview for Depression: development, reliability and validity. *Journal of affective disorders* 9, 1 (1985), 85–96.
- [54] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

- [55] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 873–883.
- [56] 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*. PMLR, 28492–28518.
- [57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [58] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*. 3–12.
- [59] Morteza Rohanian, Julian Hough, Matthew Purver, et al. 2019. Detecting Depression with Word-Level Multimodal Fusion. In *Interspeech*. 1443–1447.
- [60] Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950* (2023).
- [61] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18, 8 (2004), 1121–1133.
- [62] Guramritpal Singh Saggua, Keshav Gupta, KV Arya, and Ciro Rodriguez Rodriguez. 2022. DepressNet: A Multimodal Hierarchical Attention Mechanism approach for Depression Detection. *Int. J. Eng. Sci.* 15, 1 (2022), 24–32.
- [63] Sara Sardari, Bahareh Nakisa, Mohammed Naim Rastgo, and Peter Eklund. 2022. Audio based depression detection using Convolutional Autoencoder. *Expert Systems with Applications* 189 (2022), 116076.
- [64] Aharon Satt, Shai Rozenberg, Ron Hoory, et al. 2017. Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech*. 1089–1093.
- [65] Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 6247–6251.
- [66] Katie M Smith, Perry F Renshaw, and John Billelo. 2013. The diagnosis of depression: current and emerging methods. *Comprehensive psychiatry* 54, 1 (2013).
- [67] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.
- [68] Amanda L Stuart, Julie A Pasco, Felice N Jacka, Sharon L Brennan, Michael Berk, and Lana J Williams. 2014. Comparison of self-report and structured clinical interview in the identification of depression. *Comprehensive psychiatry* 55, 4 (2014), 866–869.
- [69] Douglas Sturim, Pedro A Torres-Carrasquillo, Thomas F Quatieri, Nicolas Malyska, and Alan McCree. 2011. Automatic detection of depression in speech using gaussian mixture modeling with factor analysis. In *Twelfth Annual Conference of the International Speech Communication Association*.
- [70] Bo Sun, Yinghui Zhang, Jun He, Lejun Yu, Qihua Xu, Dongliang Li, and Zhaoying Wang. 2017. A random forest regression method with selected-text feature for depression assessment. In *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*. 61–68.
- [71] Benedikte Timbremont, Caroline Braet, and Laura Dreessen. 2004. Assessing depression in youth: relation between the Children’s Depression Inventory and a structured interview. *Journal of Clinical Child and Adolescent Psychology* 33, 1 (2004), 149–157.
- [72] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [73] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.
- [74] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*. 3–10.
- [75] Klára Vicsi, Dávid Sztahó, and Gábor Kiss. 2012. Examination of the sensitivity of acoustic-phonetic parameters of speech to depression. In *2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 511–515.
- [76] Jingying Wang, Lei Zhang, Tianli Liu, Wei Pan, Bin Hu, and Tingshao Zhu. 2019. Acoustic differences between healthy and depressed people: a cross-situation study. *BMC psychiatry* 19 (2019), 1–12.
- [77] Yongqiang Wang, Abdelrahman Mohamed, Due Le, Chunxi Liu, Alex Xiao, Jay Mahadeokar, Hongzhao Huang, Andros Tjandra, Xiaohui Zhang, Frank Zhang, et al. 2020. Transformer-based acoustic modeling for hybrid speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6874–6878.
- [78] Peter Waxer. 1974. Nonverbal cues for depression. *Journal of Abnormal Psychology* 83, 3 (1974), 319.
- [79] Laurence White and Hannah Grimes. 2022. Articulation rate in psychotherapeutic dialogues for depression: patients and therapists. *depression* 5 (2022), 23.
- [80] James R Williamson, Elizabeth Godoy, Miriam Cha, Adrienne Schwarzenruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. 2016. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. 11–18.
- [81] Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. 2020. Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews. *arXiv preprint arXiv:2006.08336* (2020).
- [82] Uma Yadav and Ashish K Sharma. 2021. Review on automated depression detection from audio visual clue using sentiment analysis. In *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 1462–1467.
- [83] Uma Yadav and Ashish K Sharma. 2023. A novel automated depression detection technique using text transcript. *International Journal of Imaging Systems and Technology* 33, 1 (2023), 108–122.
- [84] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2020. Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences. *arXiv preprint arXiv:2010.11985* (2020).
- [85] Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. 2012. Detecting depression severity from vocal prosody. *IEEE transactions on affective computing* 4, 2 (2012).
- [86] Jiayu Ye, Yanhong Yu, Qingxiang Wang, Wentao Li, Hu Liang, Yunshao Zheng, and Gang Fu. 2021. Multi-modal depression detection based on emotional audio and evaluation text. *Journal of Affective Disorders* 295 (2021), 904–913.
- [87] Shi Yin, Cong Liang, Heyan Ding, and Shangfei Wang. 2019. A multi-modal hierarchical recurrent neural network for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 65–71.
- [88] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal Vlog Dataset for Depression Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12226–12234.
- [89] J Gerald Young, JOHN D O’BRIEN, Elane M Gutterman, and Patricia Cohen. 1987. Research on the clinical interview. *Journal of the American Academy of Child & Adolescent Psychiatry* 26, 5 (1987), 613–620.
- [90] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).
- [91] 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*. 135–143.
- [92] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2023. Hierarchical Convolutional Attention Network for Depression Detection on Social Media and Its Impact During Pandemic. *IEEE Journal of Biomedical and Health Informatics* (2023).