Multimodal Depression Detection Varies with Cross-Dataset Differences

Maneesh Bilalpur
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
mab623@pitt.edu

Saurabh Hinduja
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
sah273@pitt.edu

Sonish Sivarajkumar
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
sonishsivarajkumar@pitt.edu

Yanshan Wang
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
yanshan.wang@pitt.edu

Nicholas Allen
Center for Digital Mental Health
University of Oregon
Eugene, Oregon, USA
nallen3@uoregon.edu

Itir Onal Ertugrul
Department of Information and Computing Sciences
Utrecht University
Utrecht, Netherlands
i.onalertugrul@uu.nl

Jeff Cohn
Deliberate AI
New York, New York, USA
jeffcohn@pitt.edu

Abstract

Past work in multimodal depression detection has focused almost entirely on a single dataset with limited socioeconomic diversity and lacking in clinically validated depression diagnoses. To address these limitations in multimodal depression detection, we used clinically-validated depression diagnoses in two distinct datasets that differed by socioeconomic status (SES): A problem-solving interaction and a clinical interview. Participants in the former were of low SES; participants in the latter were of low- to middle SES status. Both included members of racial minorities. Because findings may vary with choice of modeling approach, we studied both a classical machine learning approach and two deep-learning approaches. The former consisted of handcrafted features with the Support Vector Machine classifier; the latter two were deep approaches with learnt features. One was a baseline deep approach that used a multimodal transformer with pairwise crossmodal attention. The other was our proposed multimodal transformer that used the novel Learnable Multimodal Query Crossmodal Attention (Learnable Multimodal Query CA). For each, we assessed accuracy, demographic fairness, and cross-dataset generalizability at depression detection. For both datasets, optimal embedding representations varied. The accuracy of the proposed multimodal transformer with the Learnable Multimodal Query CA and the classical approach was relatively higher. In the clinical interview, accuracy was lower for the baseline deep multimodal transformer. Generalizability across datasets was lower for both classical and deep approaches.

ICMI Companion '26, Napoli, Italy

ACM ISBN 979-8-4007-2319-3/26/10
https://doi.org/10.1145/3776591.3836736

CCS Concepts

• **Applied computing** → **Psychology**; • **Computing methodologies** → *Neural networks*; • **Human-centered computing**;

Keywords

Depression Detection, Multimodal Transformers, Fairness in Machine Learning, Cross-Dataset Analysis

ACM Reference Format:
Maneesh Bilalpur, Saurabh Hinduja, Sonish Sivarajkumar, Yanshan Wang, Nicholas Allen, Itir Onal Ertugrul, and Jeff Cohn. 2026. Multimodal Depression Detection Varies with Cross-Dataset Differences. In *Companion of the INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI Companion '26), October 05–09, 2026, Napoli, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3776591.3836736

1 Introduction

The socioeconomic burden of depression is well-documented. Depression leads to an increased risk for suicide [14] and has been found to contribute to an economic burden of a quarter of a trillion dollars in healthcare and related costs [29]. The scale of its impact highlights the need for computational approaches for depression detection using objective measures of behavior to deliver scalable and reliable assessments.

Previous work [5, 18, 31, 46, 56] has used audio, vision, and language (transcribed speech represented as text) modalities for their efficacy at depression detection. However, the majority of work is limited to the same database[7]. This lack of diversity results in a limited understanding of depression detection.

We address the lack of diversity in multimodal depression detection and the lack of clinically validated diagnosis. All participants met DSM criteria for previous depressions, and their symptom severity was assessed using standardized instruments. One dataset involved a mother-child problem-solving interaction (referred to as *problem-solving* dataset). The other dataset was a patient-clinician

interview (referred to as *clinical interview* dataset). We studied the cross-dataset effects on depression using both a classical machine learning approach and two deep approaches. The classical approach used domain-knowledge-driven multimodal handcrafted features with the Support Vector Machine (SVM) classifier. SVM was chosen because of its ability to handle small high-dimensional datasets, commonly observed in depression detection [20]. For deep approaches, we used a baseline multimodal transformer and proposed a novel paradigm to learn crossmodal dependencies. Multimodal transformers have been found to outperform other deep learning approaches [44, 46, 48, 57]. Our proposed multimodal transformer model uses a novel *Learnable Multimodal Query Crossmodal Attention* (Learnable Multimodal Query CA) using learnt features from pretrained models such as FMAE-IAT [39] for vision, WavLM [15] for audio, and LLaMA for text [47].

For unimodal insights, the modality-specific deep features with a Multi-Layer Perceptron (MLP) classifier were compared with the classical approach with constituent channels in their respective modalities (e.g., face and head dynamics, and action units channels are compared with the deep vision modality).

The growing evidence for bias in machine learning predictions resulted in increased attention to model fairness. Despite its popularity in NLP [13, 23, 26] and Computer Vision domains [12, 21, 33], it is less explored in depression detection. To address this limitation, our criteria, in addition to accuracy, include demographic fairness for race and sex of low SES participants in problem-solving dataset and low- to middle SES participants in clinical interview dataset. Beyond within dataset comparisons for accuracy and fairness, we evaluated the cross-dataset generalizability of the models by training them in one dataset to test in the unseen dataset.

Our findings on cross-dataset effects are summarized below.

(1) Optimal embeddings varied by dataset. For audio modality, earlier layers were optimal in the problem-solving dataset, and later layers were optimal in the clinical interview dataset. The optimality for the vision modality was the opposite.
(2) The multimodal deep and classical approaches were relatively more accurate in both datasets. Our Learnable Multimodal Query CA was relatively more accurate in both datasets when compared with the baseline pairwise crossmodal attention.
(3) For both multimodal deep- and classical approaches, generalizability was low between datasets.
(4) In unimodal comparisons, voice prosody (audio modality) was the strongest predictor of depression in the problem-solving dataset, but speech (text modality) was the strongest predictor in the clinical interview dataset.
(5) Further unimodal comparisons for fairness found that the classical approach with speech features (LIWC) had low fairness for race in the problem-solving dataset. More non-white women were incorrectly detected as depressed, while more white women were incorrectly detected as non-depressed. The same approach had greater fairness in the clinical interview dataset.

2 Related work

Depression detection has undergone rapid development since the early work of Cohn et al. [18] with action units and prosody. Joshi et al. [34] studied Bag-of-Words representation of spatio-temporal interest points of upper body, facial movement, and audio features for depression detection. Their audio-visual representation outperformed unimodal depression detection. Dibeklioğlu et al. [22] used dynamics of head motion and face together with switching pause and vocal pitch for depression severity prediction using the Improved Fisher Vector encoding. More recently Alghowinem et al. [5] studied feature selection over facial expressions, head motion, voice prosody, eye activity, and speech behavior. They found that the joint usage of multiple modalities improves detection performance, and an aggregate of 9 features generalize for depression detection across three distinct datasets that differ in terms of language, culture, and interaction context. Speech behavior was the strongest predictor of depression.

Xue et al. [52] used audio-text modalities for depression detection on DAIC-WoZ. They constructed audio features to contain both handcrafted-LLDs as well as embeddings from spectrogram and raw audio-signal derived using pretrained deep models such as NetVLAD [6] and wav2vec2 [8], respectively. A hierarchy of attention was used for the final detection task. Initially, multiple representations of audio modality were combined using attention. The resulting audio features were augmented with the text embeddings from BERT using a second-stage attention module to detect depression. Teng et al. [46] designed emotion prompts from transcriptions of human-agent interviews and used LLMs to identify the emotional impact from the conversations. The emotional impact was used as an additional input when predicting depression severity along with the audio, vision, and text modalities. They found that including emotion features improves detection performance and increases adaptability between related goals like affect and depression detection. Flores et al. [24] performed depression detection by treating questions from the DAIC-WoZ human-agent interview as independent samples. They used audio-visual transformers together with their novel multimodal fusion, where separate spatial and temporal attention were applied over unimodal representations. They found that questions types with longer responses (related to job, general mood, etc.) resulted in better performance than shorter response types (related to regret, depression, PTSD, etc.). The recent work by Zhang et al. [55] explored LLMs for multimodal depression detection by incorporating audio landmarks in the primarily text-based LLM. An important contribution was the discrete token representation of continuous audio modality. They used acoustic landmark representation to capture speech production events as tokens. They found that augmenting audio landmarks to the LLaMA 13B model improves detection over the native text-only model.

A large chunk of these works detect depression in a single publicly available database [7]. Consequently, it has limited our understanding of cross-dataset effects and its generalizability on depression detection. To address these concerns, this work studies depression detection in two datasets: a problem-solving dataset and a clinical interview dataset. While growing literature acknowledges the need for studying fairness in depression detection [16, 17], our work expands this focus by analyzing the fairness of classical

and deep approaches in low- to middle SES participants in both unimodal and multimodal settings.

3 Datasets

The mother-child dataset is from the Transitions in Parenting of Teens (TPOT) database [38, 43]. Groups (depressed and non-depressed) were defined based by the mother's depression status. The depressed group were 73 mothers with a history of treatment for depression (DSM-IV)[9] and current or recent elevation in depressive symptoms (PHQ-8 > 10) [36] and a non-depressed group of 75 mothers with no lifetime history of treatment for depression and no more than mild depressive symptoms (PHQ-8 < 8). The time between the assessment of current depression and the problem-solving interaction varied from same day to several weeks. The problem-solving interaction consisted of a task in which mother and child were asked to identify and resolve a problem from an updated Issues Checklist [42]. This interaction lasted about 15-minutes. All participants were of low SES and qualified for Medicaid. Eighty-five percent were White. Fifteen percent were members of one or more minority groups (American Indian/Alaskan, Native Hawaiian/Pacific Islander, African American). The children were in early adolescence. Audio was captured at 16kHz and video at 30fps and 720p resolution with dedicated hardware for mother and child.

The patient-clinician dataset is from the University of Pittsburgh depression database (Pitt) [18, 25, 28, 53]. Patients undergoing treatment for clinically diagnosed depression (DSM-IV) were assessed in a clinical interview (Hamilton Rating Scale for Depression, HRSD [30]) for depression severity. Fifty-seven depressed participants were interviewed for depression severity on up to four occasions at 1, 7, 13, and 21 weeks post-diagnosis by clinical interviewers (all female). Depression was defined as an HRSD score of 15 or higher [18]. A limitation of this archival database is its relative lack of socioeconomic description of the participants. To our knowledge they included both low- and middle SES participants. Three cameras captured the patient, and a dedicated camera recorded the interviewer. Audio was recorded using dedicated lavalier microphones for the patient and the interviewer. Only the video from the frontal-facing camera of the patient was used. Fifty participants with valid audio-visual recordings and transcriptions are used in this work. For consistency with existing work [5, 22, 35], all 135 sessions from the fifty participants were treated as independent observations for depression detection.

The segmentation of utterances and their transcription was performed manually for both databases. Each utterance was identified by its speaker along with the start and stop timestamps. For comparability between databases, the focal person for depression detection was limited to mothers in TPOT and patients in Pitt. The race and sex distributions are presented as Table 1. The prevalence of racial diversity in the databases resulted in a pronounced skew between the majority white and the remaining groups. To obtain a better sample size for the fairness analysis, we group non-majority races into a non-white group. The non-white group consists of races American Indian, Native Hawaiian, African American, Asian, and multiple races. Note that sessions with unknown race information were excluded from the fairness analysis. Unlike the problem-solving dataset, the participants in the clinical interview dataset were interviewed at regular intervals during their treatment. Hence, the participants:sessions ratio was variable. See the supplementary material for statistics related to number of sessions, utterances, and turns for depression detection in both databases.

Table 1: Participant demographics as percentage of sessions.

Demographic	Subgroup	Mother-Child	Patient-Clinician
N sessions		148	135
Race	White	85.13%	88.15%
	Non-white	13.51%	11.85%
	Unknown	1.35%	-
Sex	Men	-	35.56%
	Women	100%	64.44%

4 Methods

Our approaches include a classical approach using handcrafted features an SVM classifier and two deep approaches.

4.1 The deep approaches

The two deep approaches use embedding representations from pretrained models such as WavLM [15], FMAE-IAT [39], and LLaMA [47]. These pretrained models were chosen based on their relevance to depression detection [50, 55] or closely related affective computing tasks, such as action unit detection [39]. As the choice of embedding representations was found to affect accuracy of depression detection [50], we first assess the cross-dataset effects on embedding representations to identify optimal embeddings on the validation set. These optimal embeddings were used to train two multimodal transformer based deep approaches– the baseline multimodal transformer, which uses the pairwise crossmodal attention, and the proposed learnable multimodal query crossmodal attention.

The multimodal transformers for depression detection fuse audio, vision, and text using crossmodal attention. We use video frame embeddings from the pretrained FMAE-IAT-base model averaged over a given turn as vision modality features. Similarly, for the audio modality, we use audio frame embeddings from the pretrained WavLM-base-plus averaged over a turn. For the text modality, the embeddings of LLaMA 13B averaged over all tokens from the turn formed the features.

4.1.1 Optimal Embeddings. Embeddings of pretrained foundation models were found to exhibit optimality for downstream tasks at different layers. Wu et al. [50] found that later layers are optimal for depression detection, and earlier layers were optimal for emotion recognition. However, it is unknown if such sensitivity is observable for the same task across datasets in vision modality. We address this question using linear probing of layer-wise embeddings from the WavLM audio model and FMAE-IAT vision model in two datasets. A logistic regression model was trained with layer-wise embeddings to observe optimal performance on the validation set. The resulting optimal audio-visual embeddings were used for depression detection.

4.1.2 Baseline Multimodal Transformer. Conventional crossmodal attention [48] studied modality dependencies as an interaction between pairs of modalities. We investigate a new approach to crossmodal attention, the learnable multimodal query crossmodal attention and compare to the pairwise crossmodal attention.

Pairwise Crossmodal Attention: Pairwise crossmodal attention (Pairwise CA) captures interaction among modalities through pairwise self-attention. It adopts the standard self-attention formulation [49] to a source modality (say, audio) as the query, and the key and value share the target modality (text). Consider three modalities: audio (X_A), vision (X_V), and text (X_T). The pairwise crossmodal attention between audio and text is defined as equation-1. Since attention is directional, for a given m number of modalities, there are $\binom{m}{2}$ pairs of crossmodal attention. The final multimodal attention is the concatenation of all crossmodal attention pairs.

We define the Query as $Q_A = X_A W_{Q_A}$, Keys as $K_T = X_T W_{K_T}$, and Values as $V_T = X_T W_{V_T}$, where $W_{Q_A} \in \mathbb{R}^{d_A \times d_k}$, $W_{K_T} \in \mathbb{R}^{d_T \times d_k}$ and $W_{V_T} \in \mathbb{R}^{d_T \times d_v}$ are weights and $\mathbb{R}^{d_u}$ is a real vector of u dimensions. The crossmodal attention output from text T to audio A is presented as $Y_A := \mathrm{CM}_{T\to A}(X_A, X_T) \in \mathbb{R}^{T_A \times d_v}$:

$$\begin{aligned} Y_A &= \mathrm{CM}_{T\to A}(X_A, X_T) \\ &= \mathrm{softmax}\left(\frac{Q_A K_T^\top}{\sqrt{d_k}}\right) V_T \\ &= \mathrm{softmax}\left(\frac{X_A W_{Q_A} W_{K_T}^\top X_T^\top}{\sqrt{d_k}}\right) X_T W_{V_T}. \end{aligned} \tag{1}$$

4.1.3 Proposed Multimodal Transformer. Pairwise attention is limited to capturing dependencies between modalities as pairs. As more than two modalities often occur simultaneously (in our case, three modalities), we propose an alternative paradigm to the conventional crossmodal attention that relies on the holistic multimodal context over pairwise interactions to learn inter-modal dependencies.

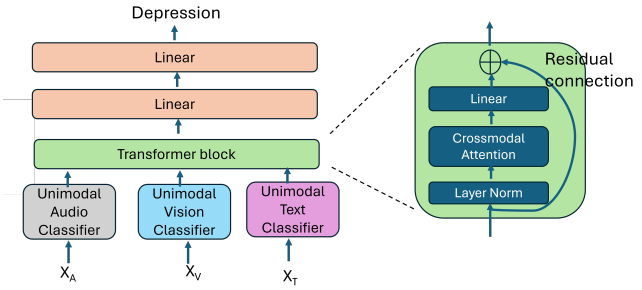

Figure 1: Multimodal Transformer model used in Depression Detection. The unimodal classifiers are two-layer MLP models. All linear layers use ReLU activation.

Learnable Multimodal Query Crossmodal Attention: The learnable multimodal query crossmodal attention addresses the limitations of the pairwise crossmodal attention by enabling trimodal interaction in crossmodal attention. To achieve this, we formulate the query to utilize multimodal context instead of one single modality. We define multimodal context in the learnable multimodal query crossmodal attention as the concatenation of embeddings from A, V, and T modalities ($X_{context} = [X_A; X_V; X_T]$). This allows each modality to simultaneously attend to all modalities. To incorporate global information, such as interactional context or other dataset-level characteristics, we also include learnable parameters ($X_L \in \mathbb{R}^{d_L}$) in the query. We use an additional feedforward layer ($W_f \in \mathbb{R}^{(d_A+d_V+d_T+d_L)\times d_q}$) to match the dimensions between the newly introduced learnable parameters and the embedding dimension of the transformer model. Equation-2 provides the mathematical formulation for learnable multimodal query crossmodal attention (learnable multimodal query CA). The novelty of our approach is unlike latent bottleneck/query-token methods [2, 37], ours does not use learnable tokens for intermediary representation; it builds an input-dependent query from all modalities and learnable parameters for direct trimodal interactions.

We use randomly initialized $d_L = 64$ dimensional learnable parameters determined based on the validation accuracy. See the supplementary for the effect of the number of learnable parameters on detection accuracy in two datasets.

$$\begin{aligned} Y_A &= \mathrm{CM}_A(X_{\text{context}}, X_A) \\ &= \mathrm{softmax}\left(\frac{Q K_A^\top}{\sqrt{d_k}}\right) V_A \\ Y_A &= \mathrm{softmax}\left(\frac{X_{\text{learnable}} W_Q {W_{K_A}}^\top X_A^\top}{\sqrt{d_k}}\right) X_A W_{V_A}, \end{aligned} \tag{2}$$

where $X_{\text{learnable}} = W_f [X_A; X_V; X_T; X_L]$.

To understand modality-level differences between the deep and classical approaches, unimodal comparisons were performed. For the unimodal deep model, a simple two-layer MLP with ReLU activation was used for depression detection. The multimodal transformer used the 512-dimensional hidden layer embeddings from the first layer outputs of the unimodal MLP layers to perform crossmodal attention between audio, vision, and text modalities. The crossmodal attention outputs for each modality were concatenated to detect depression through two linear layers with ReLU activation. Figure-1 shows the graphical overview of the multimodal transformer model. Note that unimodal models were trained separately and kept frozen when training the multimodal transformer to limit overfitting.

4.2 The classical approach

The classical approach used handcrafted features that span vision, audio, and text. These features were formulated from existing work [5, 10, 11] to capture action units (AU), face and head dynamics (FHD), speech behavior (SB), voice prosody, and linguistic behavior (LIWC) channels. The 883 resulting summary statistics features were used to train a Support Vector Machine (SVM) classifier for depression detection with the classical approach. Details about the summary statistics are in the supplementary material. While the deep approach was trained with turn-level features, the classical approach was trained with session-level features, as observed in the literature[3–5, 10, 11, 20].

4.3 Fairness: definition and metrics

A fair classifier is one that justly classifies depressed and non-depressed classes without overpredicting or underpredicting a particular demographic group. To quantify classifier fairness, the Equalized Odds Ratio (EOR) metric was used. The EOR, as shown in

Table 2: Overview of differences between the datasets

Dataset characteristics	Problem-Solving	Clinical Interview
Focal person	Mothers	Clinical trial participants
Participant sex	Women-only	Women and men
Task	Problem-solving	Depression severity assessment
Depression definition	History of depression and current symptoms	Current depression
Assessment & interaction lag	Variable lag between depression assessment and behavior sample	No lag
Recordings	Off-line-of-sight camera	Near line-of-sight camera

Equation (3), accounts for bias in both True Positive Rate (TPR) and False Positive Rate (FPR) across demographic groups in the prediction.

$$
\begin{aligned}
r_{\text{TPR}} &= \frac{\min(\text{TPR}_{\text{white}}, \text{TPR}_{\text{non-white}})}{\max(\text{TPR}_{\text{white}}, \text{TPR}_{\text{non-white}})} \\
r_{\text{FPR}} &= \frac{\min(\text{FPR}_{\text{white}}, \text{FPR}_{\text{non-white}})}{\max(\text{FPR}_{\text{white}}, \text{FPR}_{\text{non-white}})} \\
\text{EOR} &= \min(r_{\text{TPR}}, r_{\text{FPR}})
\end{aligned}
\tag{3}
$$

4.4 Cross-dataset generalization

The cross-dataset generalization experiments evaluate the robustness of classical and deep learning approaches to depression detection across the two datasets. The wide range of differences in the datasets (see Table 2) pose a significant generalization challenge to the depression detection problem. Our generalization experiments involve the model trained on the problem-solving dataset to test on the clinical interview dataset and vice-versa. Both classical and deep learning approaches (unimodal and multimodal) were assessed for generalization separately.

4.5 Cross-validation

The cross-validation setup includes a subject-independent 5-fold cross-validation to ensure that subjects in the train set did not appear in the test set. All models were trained and evaluated on the same cross-validation setup. The neural network models were optimized using the Adam optimizer [1] together with class weighting in the cross-entropy loss function to mitigate the imbalance in the datasets. The handcrafted features were normalized using the min-max normalization determined on the training set. The SVM hyperparameters for the classical approach were chosen through a grid search on the validation set for the choice of kernel and cost of misclassification. Similarly, optimal embeddings, and number of learnable parameters in the transformer model were tuned on the validation sets of the source dataset to ensure the sanctity of the test set. Due to the superior performance of early fusion over late fusion in initial experiments with the classical approach, the multimodal classical approach refers to the early fusion of features.

Although the training and evaluation for the deep approach is performed on individual turns of a multi-turn interaction, the depression label is a session-level attribute. Hence, we aggregate turn-level predictions to extract session-level predictions. A session is predicted to be depressed if the majority (50% or more) of turns are predicted to be depressed. All comparisons in this study are based on session-level outcomes. As noted above, to adjust for imbalance in the datasets, balanced accuracy (also called average recall) was reported along with Positive Agreement (PA), and Negative Agreement (NA) metrics. Note that PA is equivalent to F1-score in a two-class classification problem. The formulae for PA and NA can be found in [27]. For notational convenience, we refer to balanced accuracy as accuracy.

5 Results

5.1 Cross-dataset effect on optimal embeddings

Table-4 presents the validation results demonstrating the sensitivity of various layer embeddings to depression detection dataset. Both the problem-solving dataset and the clinical interview dataset had a difference of up to 5% between the most and the least optimal embedding layers. We found that the embeddings from the second layer of the WavLM were optimal for depression detection in the problem-solving dataset. In the clinical interview dataset, the penultimate 11^{th} layer was optimal. The FMAE-IAT vision embeddings also had a similar trend. In the problem-solving dataset, we notice that the sixth layer is optimal, while the second layer is optimal for the clinical interview dataset.

5.2 Cross-dataset effect on depression detection

The depression detection performance of multimodal attention-based fusion strategies and their comparison with the classical approach is presented as Table 3. In the problem-solving dataset, the conventional pairwise CA of modalities performed the best with 0.703 accuracy. The learnable multimodal query CA was marginally less accurate (0.696) than the pairwise CA. This trend did not generalize to the clinical interview dataset. In the clinical interview dataset, crossmodal attention with learnable multimodal query parameters performed the best with 0.701 accuracy while the pairwise CA had accuracy of 0.682. Both the multimodal transformer and the classical approach were relatively more accurate in both the problem-solving and clinical interview datasets. To understand the efficacy of including learnable parameters in the learnable multimodal query CA, we also studied the multimodal query CA that has no learnable parameters. Results in the supplementary material suggest that the learnable parameters improve performance. For further experiments on fairness analysis and generalizability, we used the multimodal transformer with the learnable multimodal query crossmodal attention fusion strategy as our choice of the

Table 3: Depression detection using deep multimodal transformers and the classical approach in Problem-Solving and Clinical Interview datasets.

		Problem-Solving			Clinical Interview		
Approach	**Fusion Strategy**	**ACC**	**PA**	**NA**	**ACC**	**PA**	**NA**
Deep (Multimodal Transformer)	Pairwise CA	**0.703**	**0.694**	**0.708**	0.682	0.644	0.674
	Learnable Multimodal Query CA	0.696	0.674	0.713	**0.701**	**0.676**	**0.676**
Classical	Early Fusion	0.703	0.675	0.724	0.696	0.678	0.677

Table 4: Validation set accuracy to determine the optimal choice of embedding layer in problem-solving dataset and clinical interview dataset for the audio and vision modalities. Numbers in bold are the best embeddings.

Embedding layer	Audio (WavLM)		Vision (FMAE-IAT)	
	Problem-Solving	Clinical Interview	Problem-Solving	Clinical Interview
1	0.732	0.500	0.608	0.571
2	**0.737**	0.500	0.649	**0.605**
3	0.723	0.519	0.611	0.600
4	0.733	0.513	0.621	0.588
5	0.729	0.501	0.673	0.591
6	0.710	0.505	**0.683**	0.576
7	0.695	0.531	0.639	0.552
8	0.691	0.531	0.606	0.500
9	0.689	0.542	-	-
10	0.690	0.550	-	-
11	0.680	**0.552**	-	-
12	0.682	0.537	-	-

multimodal transformer, as it performed well in both the problem-solving and clinical interview datasets.

We also compared the learnable multimodal query CA with existing multimodal works in both datasets. In problem-solving dataset Bilalpur et al. [10] reported 0.717 detection accuracy with AU, FHD, SB, and LIWC features and SVM classifier, in comparison our transformer approach had 0.696 accuracy. In clinical interview dataset, Kacem et al. [35] using face and head movements, and Dibeklioğlu et al. [22] using face and head movement with voice features reported 0.708 and 0.786 accuracy for three-level severity prediction respectively. We achieved 0.701 accuracy at depression detection. Since the approaches also differed by the cross-validation framework, the comparisons should be interpreted accordingly.

In unimodal experiments, both classical and deep approaches can detect depression substantially above the chance-level (0.5 accuracy) in the problem-solving and clinical interview datasets as presented in Table 6. In the problem-solving dataset, the classical approach with prosody features was most accurate at detecting depression with 0.669 accuracy. The deep alternative in audio modality detected depression with 0.648 accuracy. In the vision modality, the classical approach with AUs and the deep approach performed closely with 0.603 and 0.597 accuracy, respectively. The FHD features with the classical approach failed to generalize to the test set. In the text modality, the classical approach with LIWC features detected with 0.591 accuracy, while the accuracy for the deep approach dropped to 0.538. In the clinical interview dataset, the text modality was most effective at detecting depression. Both classical approach with LIWC features and the deep approach performed comparably with 0.665 and 0.667 accuracy respectively. In the vision modality, the classical approach using AU or FHD features was found to detect depression with similar accuracy, 0.625 and 0.621, respectively. However, the deep approach accuracy drops to 0.583. In audio modality, the deep approach (0.623) is better than the classical approach in detecting depression with prosody (0.578) or SB (0.600) features.

When unimodal and multimodal experiments for depression detection were compared, it was observed that multimodal detection was better than unimodal detection in both the problem-solving (accuracy: 0.703 multimodal; 0.669 unimodal) and the clinical interview (accuracy: 0.701 multimodal; 0.667 unimodal) datasets.

5.3 Cross-dataset effect on fairness

The Equalized Odds Ratio (EOR) for fairness is presented in Table 5. In the multimodal setting, the deep multimodal transformer achieved higher fairness than the classical multimodal approach in the problem-solving dataset for race (0.846 vs. 0.751). In the clinical interview dataset, the multimodal transformer and the classical approach had similar fairness for race (0.646 vs. 0.654), while the multimodal transformer showed higher fairness for sex (0.889 vs. 0.752). Overall, compared to the classical approach, the deep multimodal transformer had high fairness in the problem-solving dataset and comparable or higher fairness in the clinical interview dataset.

In the unimodal setting, fairness varied across modalities, features, and demographic variables. In the problem-solving dataset for race, the highest EOR was obtained by the deep text model (0.920), followed by prosody (0.823), deep vision (0.727), FHD (0.680), AU (0.646), deep audio (0.626), SB (0.615), and LIWC (0.263). In the clinical interview dataset for race, the highest fairness was obtained by deep audio (0.872) and AU (0.873), followed by deep text (0.615), prosody (0.614), SB (0.591), LIWC (0.573), FHD (0.545), and deep vision (0.541). For sex in the clinical interview dataset, AU yielded the highest EOR (0.979), followed by deep audio (0.898), SB (0.857), deep text (0.800), FHD (0.796), deep vision (0.741), LIWC (0.694), and prosody (0.651). Among the unimodal experiments, the lowest EOR for race was found with the LIWC features in the problem-solving dataset. Furthermore, when the False Positive Rate (FPR) and the False Negative Rate (FNR) between white and non-white women were compared using Fisher's exact test, the outcome was significant. The false positive rate was significantly higher for non-white women than for white women ($FPR_{non-white}$ = 0.727, FPR_{white} = 0.111; $p < 0.01$), indicating that the classical approach in text modality with LIWC features was more likely to incorrectly detect depression in non-white women. The false negative rate was significantly higher for white women than for non-white women

Table 5: Unimodal and multimodal fairness (EOR) for race and sex for mothers in problem-solving dataset and patients in clinical interview dataset. EOR ranges from 0 (least) to 1 (most) for fairness. * denotes significant difference in FPR and FNR.

Modality	Features/Approach	Mother Race & Problem-Solving	Patient Race & Clinical Interview	Patient Sex & Clinical Interview
Audio	Deep	0.626	0.872	0.898
	Prosody	0.823	0.614	0.651
	SB	0.615	0.591	0.857
Vision	Deep	0.727	0.541	0.741
	AU	0.646	0.873	0.979
	FHD	0.680	0.545	0.796
Text	Deep	0.920	0.615	0.800
	LIWC	0.263*	0.573	0.694
Multimodal	Deep (Transformer)	0.846	0.646	0.889
	Classical	0.751	0.654	0.752

Table 6: Depression detection using unimodal deep features and handcrafted behavior channels.

		Problem-Solving			Clinical Interview		
Modality	**Features**	**ACC**	**PA**	**NA**	**ACC**	**PA**	**NA**
Audio	Deep	0.648	0.602	0.672	0.623	0.557	0.534
	Prosody	0.669	0.601	0.711	0.578	0.533	0.575
	SB	0.542	0.558	0.505	0.600	0.574	0.55
Vision	Deep	0.597	0.607	0.514	0.583	0.448	0.532
	AU	0.603	0.618	0.560	0.621	0.632	0.569
	FHD	0.470	0.366	0.525	0.625	0.552	0.671
Text	Deep	0.538	0.643	0.322	0.667	0.667	0.574
	LIWC	0.591	0.458	0.665	0.665	0.653	0.645

(FNR_{white} = 0.578, $FNR_{non-white}$ = 0.111; $p < 0.05$), indicating that the classical approach in text modality with LIWC features was much more likely to miss cases of depression in white women. This observation could not be made in the clinical interview dataset despite a similar racial distribution (see Table 1).

5.4 Cross-dataset generalizability

Table 7: Multimodal cross-dataset generalization performance. Clinical Interview → Problem-Solving denotes model trained on clinical interview dataset is tested on problem-solving dataset.

Experiment	Fusion Strategy	ACC	PA	NA
Clinical Interview → Problem-Solving	Learnable Multimodal Query CA	0.500	0.527	0.133
	Early Fusion	0.500	0.527	0.133
Problem-Solving → Clinical Interview	Learnable Multimodal Query CA	0.498	0.310	0.578
	Early Fusion	0.500	0.424	0.315

Multimodal models failed to generalize across datasets as presented in Table-7. Models trained in the clinical interview dataset and tested in the problem-solving dataset performed at the chance-level. Both the deep multimodal transformer and the classical approach performed at ACC=0.500, PA=0.527, and NA=0.133. The low NA suggests that predicting the non-depressed class was more challenging than the depressed class. The generalizability from the problem-solving dataset to the clinical interview dataset also had similar outcomes, both the deep multimodal transformer (ACC=0.498, PA=0.310, and NA=0.578) and the classical approach (ACC=0.500, PA=0.424, and NA=0.315) performed similarly. They failed to generalize from the problem-solving dataset to the clinical interview.

Table 8: Unimodal cross-dataset generalization performance. Problem-Solving→Clinical Interview denotes model trained on problem-solving dataset is tested on clinical interview dataset.

		Problem-Solving → Clinical Interview			Clinical Interview → Problem-Solving		
Modality	**Features**	**ACC**	**PA**	**NA**	**ACC**	**PA**	**NA**
Audio	Deep	0.491	0.029	0.674	0.512	0.527	0.210
	Prosody	0.536	0.166	0.689	0.500	0.527	0.133
	SB	0.500	0.000	0.685	0.500	0.394	0.293
Vision	Deep	0.452	0.470	0.226	0.477	0.486	0.231
	AU	0.536	0.166	0.689	0.544	0.381	0.559
	FHD	0.500	0.097	0.522	0.500	0.660	0.000
Text	Deep	0.474	0.449	0.384	0.500	0.660	0.000
	LIWC	0.561	0.653	0.117	0.523	0.596	0.340

The cross-dataset generalizability of unimodal experiments between problem-solving dataset and clinical interview dataset is provided in Table 8. Among the models trained in the problem-solving dataset and tested in the clinical interview dataset, a significant drop in performance was noted across audio, vision, and text modalities when compared to within-dataset performance (see Table 6 and Table 3). We notice that unimodal models with the deep approach failed to generalize across datasets in all three modalities, as they were less accurate than a chance-level accuracy of 0.500. Although the classical approach with individual channels outperformed the unimodal deep approach with modest generalization accuracy, their low PA limits their validity in depression detection, as they were much better at detecting the non-depressed subjects than the depressed subjects. The generalization accuracies for LIWC (0.561), prosody (0.536), and AU (0.536) features exceeded the chance-level, the low class-level agreements suggest that they underpredict one of the two classes. We observe a similar behavior with the model trained in the clinical interview dataset and tested in the problem-solving dataset. Unimodal models with the deep

approach failed to generalize for all three modalities and the highest accuracy was obtained using audio features (0.512). However, they were only nominally better than the chance-level. Among individual channels with the classical approach, AU features had the highest accuracy (0.544) with PA (0.381), followed by LIWC features (accuracy=0.523, PA=0.596). However, the generalizability of clinical interview dataset models was also limited by the underprediction of one of the two classes.

6 Discussion

We explored multimodal modeling of depression in two datasets that overlapped in demographics and depression ascertainment. Participants in the problem-solving dataset were all female, low SES, and had a history of diagnosed depression by DSM-IV criteria and recent or current symptoms. Participants in the clinical interview were males and females, had more recent history of diagnosed depression by the same DSM-IV criteria, and their current symptom severity was assessed using a clinical interview on the same day that their multimodal behavior was obtained.

Our findings indicate varied cross-dataset effects. Optimal representations of both audio and video varied with dataset. The second layer of audio WavLM was optimal in the problem-solving dataset, while the 11th layer was optimal in the clinical interview. The 6th layer of vision FMAE-IAT was optimal in the problem-solving dataset, while the second layer was optimal in the clinical interview. Given that deep models capture a hierarchy of features with model depth[40, 41, 51, 54], these findings suggest that the depression relevant behaviors observed may differ by datasets, which has implications for future work in multimodal depression detection.

Performance depended in part on both fusion mechanism and dataset. The accuracy of baseline multimodal transformer and the classical approach were relatively higher only in the problem-solving dataset. Our proposed multimodal transformer with learnable multimodal query crossmodal attention achieved performance comparable to the classical approach in both datasets.

Despite strong within-dataset performance, inter-modal dependencies learned in one setting failed to transfer reliably to the other. This aligns with the substantial differences between datasets, with the understanding that differences in demographics or related factors may have played a role.

In multimodal comparisons, the deep approach was fairer than the classical approach in both datasets; while fairness in individual modalities varied by dataset. The classical approach with LIWC features exhibited low fairness for mother's race in the problem-solving dataset but not in the clinical interview. This implies that multimodal fusion may mitigate some demographic disparities. The same was not found in unimodal settings. Consequently, any deployment-facing solution should also consider auditing fairness.

Unimodal results for depression detection also offered fine-grained insights on cross-dataset effects, for the strongest predictor modality In the problem-solving dataset, the classical approach with prosody features from audio modality was most accurate, whereas in the clinical interview dataset, LIWC features from text was most accurate. This divergence is consistent with differences in task structure: clinical interview sessions were semi-structured HRSD interviews, where language-based symptom assessment is foundational, so LIWC's strong performance suggests that verbal responses to HRSD questions effectively encode depression-relevant signals. In contrast, the problem-solving task was unstructured, hence the more variable speech encourages other modalities to encode depression-relevant signals. The superior performance of prosody also aligns with prior evidence [19] that depression-related behaviors (e.g., psychomotor retardation) can be captured via computational measures of voice (e.g., voice quality, formants, MFCCs). Across unimodal comparisons, the deep approach did not outperform the classical approach using individual-channel features.

7 Conclusion

We examined the cross-dataset effects on multimodal depression detection. We compared a classical multimodal approach and two multimodal transformer based deep approaches. We found differences in depression detection along multiple fronts, including optimal embedding representations, detection accuracy, demographic fairness in detection, and cross-dataset generalizability.

The optimal embedding representations for the deep approaches varied by dataset for both audio and vision modalities. Our proposed multimodal transformer for depression detection and the classical approach were relatively highly accurate at depression detection. Despite a strong within-dataset accuracy, both multimodal classical and deep approaches achieved low cross-dataset generalizability.

From unimodal experiments, prosody (audio modality) was the strongest predictor in the problem-solving dataset, but speech (text modality) was the strongest predictor in the clinical interview dataset. Unimodal fairness analysis found that the classical approach had low fairness for race with text modality in the problem-solving dataset but not in the clinical interview dataset.

Three limitations may be noted. One, specific differences due to individual factors of the datasets could not be disambiguated. In future work, datasets that disambiguate individual factors in depression detection are needed. Two, a larger number of multimodal models would be informative. Three, temporal modeling may capture localized depression-related phenomena that summary statistics and average-pooled features miss. [32, 45].

8 Safe and Responsible Innovation Statement

Data were from two sources [25, 43]. Data collection protocols and analysis for each were approved by the respective institutional Ethics Review Boards. All participants gave their informed consent prior to data collection. The algorithms we developed are intended for research use. More development and testing is required to establish the clinical validity of detection approaches for the intended populations. This should include educating stakeholders, including regulatory bodies, providers and clients, about possible demographic effect on detection approaches.

Acknowledgments

This work was supported in part by the U.S. National Institutes of Health through U.S. National Institute of Mental Health award MH096951.

References

[1] Kingma DP Ba J Adam et al. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* 1412, 6 (2014).

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[3] Sharifa Alghowinem, Roland Goecke, Jeffrey F Cohn, Michael Wagner, Gordon Parker, and Michael Breakspear. 2015. Cross-cultural detection of depression from nonverbal behaviour. In *2015 11th IEEE International conference and workshops on automatic face and gesture recognition (FG)*, Vol. 1. IEEE, 1–8.

[4] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parker, and Michael Breakspear. 2013. Eye movement analysis for depression detection. In *2013 IEEE International Conference on Image Processing*. IEEE, 4220–4224.

[5] Sharifa Mohammed Alghowinem, Tom Gedeon, Roland Goecke, Jeffrey Cohn, and Gordon Parker. 2020. Interpretation of depression detection models via feature selection methods. *IEEE Transactions on Affective Computing* (2020).

[6] 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*. 5297–5307.

[7] Umut Arioz, Urška Smrke, Nejc Plohl, and Izidor Mlakar. 2022. Scoping review on the multimodal classification of depression and experimental study on existing multimodal models. *Diagnostics* 12, 11 (2022), 2683.

[8] 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 (2020), 12449–12460.

[9] Carl C Bell. 1994. DSM-IV: diagnostic and statistical manual of mental disorders. *Jama* 272, 10 (1994), 828–829.

[10] Maneesh Bilalpur, Saurabh Hinduja, Laura Cariola, Lisa Sheeber, Nicholas Allen, Louis-Philippe Morency, and Jeffrey F Cohn. 2023. SHAP-based Prediction of Mother's History of Depression to Understand the Influence on Child Behavior. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 537–544.

[11] Maneesh Bilalpur, Saurabh Hinduja, Laura A Cariola, Lisa B Sheeber, Nick Allen, László A Jeni, Louis-Philippe Morency, and Jeffrey F Cohn. 2023. Multimodal Feature Selection for Detecting Mothers' Depression in Dyadic Interactions with their Adolescent Offspring. (2023).

[12] Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. 2024. The dark side of dataset scaling: Evaluating racial classification in multimodal models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1229–1244.

[13] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[14] Stephanie L Burcusa and William G Iacono. 2007. Risk for recurrence in depression. *Clinical psychology review* 27, 8 (2007), 959–985.

[15] 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 (2022), 1505–1518.

[16] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. 2024. Fairrefuse: Referee-guided fusion for multi-modal causal fairness in depression detection. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

[17] Jiaee Cheong, Micol Spitale, and Hatice Gunes. 2023. "It's not Fair!"–Fairness for a Small Dataset of Multi-modal Dyadic Mental Well-being Coaching. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[18] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. 2009. Detecting depression from facial actions and vocal prosody. In *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, 1–7.

[19] Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. 2011. An investigation of depressed speech detection: Features and normalization. In *INTERSPEECH 2011 12th Annual Conference of the International Speech Communication Association*. International Speech Communication Association, 2997–3000.

[20] 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.

[21] Sepehr Dehdashtian, Ruozhen He, Yi Li, Guha Balakrishnan, Nuno Vasconcelos, Vicente Ordonez, and Vishnu Naresh Boddeti. 2024. Fairness and Bias Mitigation in Computer Vision: A Survey. *arXiv preprint arXiv:2408.02464* (2024).

[22] Hamdi Dibeklioğlu, 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.

[23] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.

[24] Ricardo Flores, ML Tlachac, Avantika Shrestha, and Elke A Rundensteiner. 2025. WavFace: A Multimodal Transformer-based Model for Depression Screening. *IEEE Journal of Biomedical and Health Informatics* (2025).

[25] Ellen Frank, Giovanni B Cassano, Paola Rucci, Wesley K Thompson, Helena C Kraemer, Andrea Fagiolini, Luca Maggi, David J Kupfer, M Katherine Shear, Patricia R Houck, et al. 2011. Predictors and moderators of time to remission of major depression with interpersonal psychotherapy and SSRI pharmacotherapy. *Psychological medicine* 41, 1 (2011), 151–162.

[26] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational linguistics* 50, 3 (2024), 1097–1179.

[27] Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, and Jeffrey F Cohn. 2017. Sayette group formation task (GFT) spontaneous facial expression database. In *FG*. IEEE, 581–588.

[28] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. 2014. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing* 32, 10 (2014), 641–647.

[29] Paul Greenberg, Abhishek Chitnis, Derek Louie, Ellison Suthoff, Shih-Yin Chen, Jessica Maitland, Patrick Gagnon-Sanschagrin, Andree-Anne Fournier, and Ronald C Kessler. 2023. The economic burden of adults with major depressive disorder in the United States (2019). *Advances in Therapy* 40, 10 (2023), 4460–4479.

[30] Max Hamilton. 1960. A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry* 23, 1 (1960), 56.

[31] Albert Haque, Michelle Guo, Adam S Miner, and Li Fei-Fei. 2018. Measuring depression symptom severity from spoken language and 3D facial expressions. *arXiv preprint arXiv:1811.08592* (2018).

[32] Saurabh Hinduja, Ali Darzi, Itir Onal Ertugrul, Nicole Provenza, Ron Gadot, Eric A Storch, Sameer A Sheth, Wayne K Goodman, and Jeffrey F Cohn. 2024. Multimodal Prediction of Obsessive-Compulsive Disorder and Comorbid Depression Severity and Energy Delivered by Deep Brain Electrodes. *IEEE transactions on affective computing* 15, 4 (2024), 2025–2041.

[33] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Quantifying societal bias amplification in image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13450–13459.

[34] Jyoti Joshi, Roland Goecke, Sharifa Alghowinem, Abhinav Dhall, Michael Wagner, Julien Epps, Gordon Parker, and Michael Breakspear. 2013. Multimodal assistive technologies for depression diagnosis and monitoring. *Journal on Multimodal User Interfaces* 7 (2013), 217–228.

[35] Anis Kacem, Zakia Hammal, Mohamed Daoudi, and Jeffrey Cohn. 2018. Detecting depression severity by interpretable representations of motion dynamics. In *FG*. IEEE, 739–745.

[36] 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 (2009), 163–173.

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[38] Benjamin W Nelson, Lisa Sheeber, Jennifer H Pfeifer, and Nicholas B Allen. 2021. Affective and Autonomic Reactivity During Parent–Child Interactions in Depressed and Non-Depressed Mothers and Their Adolescent Offspring. *Research on Child and Adolescent Psychopathology* 49, 11 (2021), 1513–1526.

[39] Mang Ning, Albert Ali Salah, and Itir Onal Ertugrul. 2025. Revisiting Representation Learning and Identity Adversarial Training for Facial Behavior Understanding. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–10.

[40] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature Visualization. *Distill* (2017). doi:10.23915/distill.00007 https://distill.pub/2017/feature-visualization.

[41] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 914–921.

[42] Ronald J Prinz, Sharon Foster, Ronald N Kent, and K Daniel O'Leary. 1979. Multivariate assessment of conflict in distressed and nondistressed mother-adolescent dyads. *Journal of applied behavior analysis* 12, 4 (1979), 691–700.

[43] Lisa Sheeber, Jessica Lougheed, Tom Hollenstein, Craig Leve, Kavya Mudiam, Catherine Diercks, and Nicholas Allen. 2023. Maternal aggressive behavior in interactions with adolescent offspring: Proximal social–cognitive predictors in depressed and nondepressed mothers. *Journal of psychopathology and clinical science* (2023).

[44] Hao Sun, Yen-Wei Chen, and Lanfen Lin. 2022. Tensorformer: A tensor-based multimodal transformer for multimodal sentiment analysis and depression detection. *IEEE transactions on affective computing* 14, 4 (2022), 2776–2786.

[45] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology* (2025).

[46] Shiyu Teng, Shurong Chai, Jiaqing Liu, Tomoko Tateyama, Lanfen Lin, and Yen-Wei Chen. 2024. A sentiment pre-trained text-guided multimodal cross-attention transformer for improved depression detection. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1–4.

[47] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[48] 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. 6558.

[49] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[50] 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)*. IEEE, 1–5.

[51] Yaoxun Xu, Shi-Xiong Zhang, Jianwei Yu, Zhiyong Wu, and Dong Yu. 2024. Comparing discrete and continuous space llms for speech recognition. *arXiv preprint arXiv:2409.00800* (2024).

[52] 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)*. IEEE, 6790–6794.

[53] Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. 2012. Detecting depression severity from vocal prosody. *IEEE transactions on affective computing* 4, 2 (2012), 142–150.

[54] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

[55] 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* (2024).

[56] Wenjie Zheng, Qiming Xie, Zengzhi Wang, Jianfei Yu, and Rui Xia. 2025. Towards Explainable Multimodal Depression Recognition for Clinical Interviews. *arXiv preprint arXiv:2501.16106* (2025).

[57] Jianghai Zhou, Jike Ge, Zuqin Chen, Jie Tan, and You Li. 2025. MDD-MARF: A multimodal depression detection model based on multi-level attention mechanism and residual fusion. *Journal of Biomedical Informatics* (2025), 104965.